Mission Critical Communication Systems: Performance Evaluation, Enhancement, and Modelling

by

Ashraf Ali

Faculty of Computing, Engineering, and Sciences

University of South Wales

Pontypridd, Cardiff CF37 1DL, Wales, UK

A thesis submitted in partial Fulfilment of the requirement for the degree of Doctor of Philosophy

July 2018

Dedication

To my Father Dr Abidrabbu for his encouragement, To my Mother Huda for her Endless care, To my Wife Marwa for her love and support, To my baby daughter Juana for her loud out of routine support at midnight, and finally for the Humanity and life-saving team members who use Mission Critical Systems

ABSTRACT

Mission Critical Systems (MCSs) are needed mainly as All-Time-Available backbone system for emergencies, crises, and disaster scenarios. They are used in Public Protection and Disaster Relief (PPDR) operations, Utility Networks, and Intelligent Transportation Systems. Due to the critical nature of MCSs, they need to adhere to strict requirements to ensure the accomplishment of tasks and duties that are usually strongly associated with human lives and national security. Delay, Interoperability, Availability, Reliability, Security, and Resilience are some of the requirements that need to be met by an MCS to ensure optimum functionality as required by its users. There are two main deployment options for an MCS; the first is a dedicated system, such as TErristrial Trunked Radio (TETRA), that is designed to meet all the MCS requirements and which operates only for mission critical operations and tasks, while the second is a commercial general purpose mobile communication system that can be used for both conventional mobile communication as well as an MCS. The first option can only support low data rates to support limited applications. Whereas the second option can provide users with broadband services using a more generic communication system that can be used as an MCS as well as conventional mobile communications system, but it lacks the scalability and reliability support provided by dedicated MCSs.

This research, articulated in this thesis, de-risks the compliance of a new mission critical system proposed by standardisation bodies and governmental authorities, merging the strengths of both dedicated and generic mobile communication systems, to meet the MCS requirements. Delay of Session Initiation Protocol (SIP) signalling between different entities in the core network, such as in IP Multimedia Subsystem (IMS), is considered one of the major factors that contributes to the Scalability, Reliability, and end-to-end delay challenges of generic next generation MCS. This thesis, presents a literature study of recent findings and enhancements in signalling domain, in addition to an appraisal of the multimedia communications signalling performance metrics and evaluation techniques along with a bottleneck analysis of the core network entities. A systematic methodology to run a set of experiments and evaluation techniques was developed by referring to recent methodological techniques found in the literature and a new framework was designed and evaluated to overcome the challenges of dedicated systems.

The evaluation, using simulations and experimental testbed, showed the effect of SIP signalling delay over the overall system performance in terms of scalability and robustness. Simulation results showed the delay effects introduced by the access technology over the entire system performance. Moreover, the testbed results, using different systematic scenarios, showed the core network part effect over the entire system performance especially the scalability, responsiveness, and reliability aspects. Both, the simulations and testbed experiments, show that there is a need to improve the performance of the current real-time processing techniques at the user end interface (access technology domain) and the core end subsystems (IMS and related interfaces) to compensate for the performance impairments. The framework proposed, was able to process the traffic in real time and send feedback information utilising the information found in the header of different network layers to decide a load balancing mechanism able to minimise the end-to-end delay and scalability shortcoming during heavy load scenarios.

ACKNOWLEDGMENTS

Firstly, I would like to thank the Director of the Study, Prof. Andrew Ware, who despite his late appointment as a director of study for my research in the fourth year of the project, has never hesitated to give me the time needed to read, provide hints, proofread, and encourage me to finish the research, not to forget his invitations to have a meeting while dining out. His support is outstanding and to be remembered for long.

I would like to thank Prof Khalid Al-Begain, for his support especially during the first three years of the research project. His guidance and hints has shaped the way this research has been carried out. With simple hints and conversation in the middle of his busy time schedule, I always found something useful carry out the research. I enjoyed attending his lectures, working with him in organising conferences, editing a book, and having short and lengthy talks. I would like to thank Prof Ifiok Otung as well for his support and offering his help whenever needed, although the topic was not quit related to his research field, he never hesitated to give me the needed time to listen and support me during this journey.

I would like to thank University of South Wales for giving me this remarkable experience to study PhD in Wales. The welsh tradition and hospitality will be unforgettable. I thank as well the research office, represented by Ms Llinos Spargo, who has been so helpful and supportive during the four years journey. The thanks goes as well to the Hashemite University in Jordan who sponsored my study at USW for the first four year, without their support, the dream will not turn into reality.

Last but not least, to my loving family in Jordan, my Mother Huda and my father Dr. Abidrabbu, who encouraged and supported me since I opened my eyes to this life. To my family her in Wales, My Wife Marwa and My baby daughter Juana, who with their love, patience and support gave me the power to continue until the end and beyond.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

3GPP: 3rd Generation Partnership Project.

4G: Fourth Generation Technology

5G : Fifth Generation technology

AGA: Air-Ground-Air

APCO: Association of Public-Safety Communications Officials-international

ARP: Allocation ad Retention Priority

AS: Application Server

ACK: Acknowledgment message

BPSK: Bipolar Phase Shift Keying

BER: Bit Error Rate

BLER: Block Error Rate

CDMA: Code Division Multiple Access

C-RNTI: Cell Radio Network Temporary Identifier

DQPSK: Differential Quadrature Phase Shift Keying

D2D: Device to Device Communication

DMO: Direct Mode Operation

DTLS: Datagram Transport Layer Security

DCCH : Downlink Control Channel

DL-SCH: Downlink Shared Channel

ESMCP: Emergency Services Mobile Communications Program

ETSI: European Telecommunications Standards Institute

EUTRAN: Evolved Universal Terrestrial Radio Access Network

E-UMTS: Evolved Universal Mobile Telecommunication Standard

EPC: Evolved Packet Core

Enb: Evolved Node B

EPS: Evolved Packet System

E-CSCF: Emergency Call Session Control Function

ECGI : E-UTRAN Cell Global Identifier

ENUM: Telephone Number Mapping

FDMA : Frequency Division Duplex

FEC: Forward Error Correction

FBC: FeedBack Channel

GPRS: General Packet Radio Services

GBR: Guaranteed Bit Rate

GUTI: Globally Unique Temporary Identifier

GUMMEI: Globally Unique MME Identifier

GTP: GPRS Tunnelling Protocol

HSS: Home Subscriber Station

H2H: Human to Human Communication

HNB: Home Node B

ITS: Intelligent Transport Services

IMS: IP Multimedia Subsystem

ISO: International Standardisation Organisation

IoT: Internet of Things

IoE: Internet of Everything

IETF: Internet Engineering Task Force

I-CSCF: Interrogator Call Session Control Function

IRA: Ineffective Registration Attempts

ITU: International Telecommunication Union

IPPM: IP performance Metrics

IPsec : Internet Protocol Security

IPv6 : Internet Protocol version 6

IPv4 : Internet Protocol version 4

IMSI: International Mobile Subscriber Identity

KPI: Key Performance Indicator

LTE: Long Term Evolution

LDF: Load Detection Function

MAC: Medium Access Control

MCS: Mission Critical System

MCCS: Mission Critical Communication System

MIMO: Multi Input Multi Output

MNO: Mobile Network Operator

MME: Mobile and Management Entity

M2M: Machine to Machine Communication

MAR: Multimedia Authentication Request

MAA: Multimedia Authentication Answer

MSSTOI: Mean Session Setup Time Originated from IMS

MSU: Mean Session Utilization

MGW: Multimedia Gateway

MGCF: Multimedia Gateway Control Function

MANET: Mobile Ad hoc Mobile Network

MmWave: mille-meter Wave

MIB: Master Information Block

MBR: Maximum Bit rate

MSB: Most Significant Bit

NGO: Non-Governmental Organisation

OFDMA: Orthogonal Frequency Multiple Access

PPDR: Public Protection and Defence Relief

PSAC: Public Safety Advisory Committee

PTT: Push to Talk

PS: Public Safety

PSS: Public Safety System

PMR: Professional Mobile Radio

PLMR: Professional Land Mobile Radio

P25: APCO Project number 25

PoC: Push to Talk over Cellular

PSN: Public Safety Networks

PDF: Probability Distribution Function

PDO: Packet Data Optimized

PM: Phase Modulation

P-GW: Packet Data Network Gateway

PCRF: Policy and Charging Rules Function

PD: Processing Delay

PSTN: Public Switched Telephony Network

PDN: Packet Data Network

PDP: Packet Data Protocol

PLMN: Public Land Mobile Network

PDU: Protocol Data Unit

PMN: Public Mobil Network

PRACK: Provisional Response Acknowledgment

QoS: Quality of Service

QoE : Quality of Experience

QPSK: Quadrature Phase Shift Keying

QAM: Quadrature Amplitude Modulation

QoSg: Quality of Signalling

QCI: Quality of Service Class Identifier

RTP: Real Time Protocol

RAN: Radio Access Network

RRD: Registration Request Delay

RpD: Response Delay

RpT: Responses Transmits

RRC: Radio Resource Control

RLC: Radio Link Control

RAB: Radio Access Bearer

RAR: Re-Auth-Request

RAA: Re-Authentication-Answer

RRps: Registration Requests per second

SIP: Session Initiation Protocol

SDP: Session Description Protocol

SLA: Service Level Agreement

SwMI: Switching and Management Infrastructure

S-GW: Serving Gateway

SNR: Signal to Noise Ratio

SINR: Signal to Interference Noise Ratio

SDD: Session Disconnect Delay

SRD: Session Request Delay

S-CSCF: Serving Call Session Control Function

SDT: Session Duration Time

SER: Session Establishment Ratio

SEER: Session Establishment Effectiveness Ratio

SP: Service Provider

SUT: System Under Test

SIB: System Information Block

S1AP: S1 Application Protocol

SIB1: System Information Block Type 1

SIB2: System Information Block Type 2

SIB3: System Information Block Type 3

SIB4: System Information Block Type 4

SIB5: System Information Block Type 5

SIB6: System Information Block Type 6

SIB7: System Information Block Type 7

SIB8 : System Information Block Type 8

SIB9 : System Information Block Type 9

SRB: Signalling Radio Bearer

SRPD: Service Request Processing Durability

TETRA: TErrestrial Trunked RAdio

TCCA: The Critical Communication Association

TCP: Transport Control Protocol

TDD: Time Division Duplex

TDMA: Time Division Multiple Access

TMO: Trunked Mode Operation

TLS: Transport Layer Security

TS: Technical Standard

UMTS: Universal Mobile Telecommunication System

UAR: User Authentication Request

UAA: User Authentication Answer

UE: User Element

UDP: User Datagram Packet

UUD: User to User Delay

UL-CCH: Uplink Control Channel

VoIP: Voice over IP

VoLTE: Voice over Long Term Evolution

WiMAX: Wireless interoperable Mobile Access System

WLAN: Wireless Local Area Network

WAN: Wide Area Network

Chapter 1: INTRODUCTION

1.1 INTRODUCTION

Many governmental and non-governmental organizations (NGOs) play critical roles in the operation of mission critical systems. The services provided by such organization are of a type that make them less tolerant to executional or operational errors. Such services are referred to as mission critical services. The criticality of such services implies that they have a set of special requirements that distinguish them from other services. They should be available anytime, anywhere, within the service operational scope. Moreover, they need to be able to function at the extremity of their capabilities regardless of the operational circumstances or running conditions.

In this Chapter, the definition of Mission Critical Systems, users, and services will be introduced, then main applications of such system will be listed, the general requirements will be investigated, and finally a distinction between different types of Mission Critical Systems will be briefly discussed.

'Mission Critical' is defined by TETRA (Terrestrial Trunked Radio, 2003) and The Critical Communication Association (TCCA) as a function whose failure would lead to catastrophic consequences that would place public order or public security at risk. A system that provides such critical functionality must have suitable inbuilt functionality, interoperability, security, and the wherewithal to maintain its availability. Mission critical users are those with responsibility for the welfare, health, security and safety of the public. Law enforcement forces, fire fighters, emergency and medical services, rescue services, military forces, utility staff members, and transport services members, are all example of mission critical users. The concept of a Mission Critical Communication Systems (MCCS) refers to the hardware, software and communication facilities that allow mission critical users to communicate with each other and liaise with command centres securely and dependably for the sake of providing mission critical services, wherever and whenever the services require special communication solutions. (TETRA and Critical Communication Association, 2010).

1.2 APPLICATIONS OF MISSION CRITICAL SYSTEMS

Based on the definition mentioned above, The Mission Critical Systems (MCSs) are considered as systems that provide critical services for a certain target group. MCS is needed mainly as All-Time-Available backbone system to be accessible by system users and ensuring connectivity between clients and satisfying the service requirements with acceptable service quality based on the task and service type demanded by the client. The American Public Safety Advisory Committee (PSAC) in the United States describes every system that is capable of providing: an immediate communication with instantaneous connectivity, reliability to minimize the short term disruptions, and in most cases secure in order not to be accessed by unauthorized users. Such systems are required by many applications that will be briefly introduced in the following subsections.

1.2.1 Public Protection and Disaster Relief.

Public Protection and Disaster Relief (PPDR) is one of the MCSs that is needed by law enforcement, emergency, and medical services teams during emergencies, crises, and disaster scenarios (Baldini, Karanasios, Allen, Vergari, 2014), the system is used mainly for liaison between different PPDR members to make sure that all needed resources are available during and after the crises that may occur anytime and anywhere. Mission Critical Communication System is needed for PPDR system since it provide services of critical nature that may make the difference in saving precious lives during harsh operating conditions. One of the main challenges is making the required resources available and dedicated for the mission critical communication system to provide an enhanced communication system that is scalable to high number of users along with diverse concurrent running applications.

PPDR members use the talk-group function to communicate with each other or with the control room, setting the talk-group instantaneously on site is crucial for optimum teamwork functionality. Therefore, supporting group communication is one of the most important requirements in the MCS that serve PPDR members. The group communication is done via Push-To-Talk function to address another user or set of users within a group in a half-duplex communication. This implies that a special hardware terminals must be designed and manufactured to support the PTT functions.

1.2.2 Utilities

Utilities such as Gas, Water, and Electricity is considered of great importance nowadays for country's economy and development plans. The Electricity outage, for example, has tremendous impact on all other sectors. Therefore, the Electricity Transmission Grid is considered one of the Mission Critical Systems that is supposed to provide a service or a functionality which, in case of failure, may lead to catastrophic results. What applies to Electricity also applies on the Gas services, especially that the Gas Transmission Networks are operated across multiple countries. Similarly, Water Network is also critical by nature and need to be operated by MCS to ensure the reliability of the provided service.

All the three utilities sectors need special requirements to support the provided services by each sector. Ensuring the connectivity using a reliable system, group talk between staff members on ground, and data monitoring and collection within specific time limits to be followed by automated response, are all considered examples of services that need special requirements for a system that has a critical nature.

1.2.3 Intelligent Transport Systems.

The wide application of technological operational methods in the transportation sector, leaded to what is called as Intelligent Transport Services (ITS). Normally, the number of applications in IT'S is more than that in Utilities and PPDR sectors. Most of the ITS applications require narrow band data, with other applications, such as video monitoring of traffic, require broadband connectivity to control centres. Controlling the transportation sector is considered a mission critical task that need to be operated via a MCS. However, in the ITS sector there are other entertainment services that are considered of non-critical nature such as providing

broadband connectivity for passengers, which implies that there is a need to define the requirements of the system based on different kinds of services provided.

1.3 MCS FUNCTIONAL AND OPERATIONAL REQUIREMENTS.

Public Safety (PS) communications are very important in incidents where too many emergency forces, such as police, civil defence, medical staffs, and firemen, need to liaise and collaborate to ensure dealing with critical incidents and events within strict time limits. Due to the nature of such incidents or accidents, every second counts in saving a precious life. Hence, a set of strict requirements need to be met to guarantee a reliable structure and framework that offer reliable and resilient services. Due to critical nature for such systems which distinguishes it from other commercial mobile communications, we refer to it as Professional Mobile Radio (PMR).

The requirements of public safety communication systems is considered more strict than the commercial communication networks, the requirements are governed by the following factors: the type of services provided to the end users, the technology being used, the working environment scenarios, and the specific role or the functionality of the MC users.

As mentioned in the previous section, there are three main sectors that applies Mission Critical functions; PPDR, Utility, and ITS. PPDR has a special operational and functional requirements such as the need for broadband services, voice communication, back to back communication for first responders, secure communication, etc. the aforementioned functional requirements determine the technical requirements of the MCS needed for PPDR operations as will be presented in next sections.

The Utilities also has a functional and operational requirements to ensure hassle free operation such as continuous data monitoring, automated control, system restore ability, service coverage, low metering running cost, etc. Similarly, ITS has diverse applications such as traffic signals management, vehicle detection and tracking, Closed-Circuit Television (CCTV), etc. interoperability and security over narrowband communication medium are some of MCS requirements in ITS sector.

A full description of MCS requirements will be presented in the following sections.

1.4 TYPES OF MISSION CRITICAL SYSTEMS

There are three deployment options for Mission Critical Communication Systems; the first one is the dedicated mission critical systems, which, as the name implies, is only dedicated for the mission critical communications tasks and operations. The second one is the commercial Mission Critical System that is used for mobile communications in addition to the Mission Critical operations. Finally, there is a hybrid approach that combines both the commercial and dedicated MCS0 (Ferries, et.al, 2013).

1.4.1 Dedicated MC Communication Systems

The dedicated mission critical communication is already designed to meet the strict requirements of the mission critical services and communications. TETRA in Europe and P25 in USA are the most common standards being used nowadays for the mission critical communications. We will investigate in depth the dedicated MC systems in the next sections.

1.4.2 Commercial MC Communication Systems

The commercial communication systems are considered as a general purpose system that is used for public mobile communication and also whenever needed it should be ready for being used by the PPDR clients as a mission critical communication systems. The optimum goal is to have a system that works as a replica of a dedicated mission critical communication system using commercial mobile system equipment. (Blom, 2008) shows the feasibility of having Public Safety System over commercial cellular Technology. Clearly, this introduces a lot of design and enhancement challenges and constraints that will be investigate in depth in the next sections.

1.4.3 Hybrid MC Communication Systems

Having a dedicated MCS is of high cost and expenses. Governments are thinking about having a set of technologies to be running altogether for supporting the mission critical operations. This will provide a smooth cheaper transition toward a full implemented commercial MCS. in the hybrid approach, all the technologies such as TETRA, Commercial LTE, Wi-Fi, WiMAX, Satellite communications, legacy 2G and 3G. All can work to support the mission critical operations in the three sectors.

In (Baldini, 2012) a framework that combines multiple communication technologies was presented to be considered for operation in Europe. The selection of the technology depends on the needed bandwidth and coverage along with the requirements needed by the MC sector. However, the management between different technologies is complex and require a clear division of responsibilities.

1.5 MIGRATION FROM CURRENT DEDICATED MCSS

As described before, dedicated mission critical systems are reliable and designed for the Mission Critical operations and able to provide users with low and mid-range data services. However, due to its high operational and capital expenditures without the ability to provide broadband services with only a limited functionality to voice oriented and data limited services makes it not the best option for the future mission critical communications. In this section, the emerging of mission critical data needs will be presented followed by a case study of proposals for providing mission critical services over non-dedicated MCS structures, then the feasibility of commercial mobile communications will be discussed, and finally the need for hardening the generic commercial mobile communication to be eligible for being used as a MCS will be discussed.

1.5.1 Mission Critical Users Demands

The user demands for more data due to the changes in applications and users' needs has also affected the Mission Critical Systems way of operation. In (TETRA, 2010) the trends in the use of mobile applications in the public sector shows that the public safety sector community is following the wider society for the need to access a wider set of applications due to the change in the way of working. For example, the enhancement in the effectiveness and efficiency of the incident response requires that PPDR members access a simultaneous set of applications. A more need for data services along with voice services is also essential managing planned and unplanned major events. Nowadays PPDR operations are increasingly need more bandwidth to support the mixture of different multimedia services that support their daily operations, transmitting videos, images; high definition audio are becoming as much important as conventional voice communications.

Video conference call between the ambulance and the hospital, sending a crime scene photos and diagrams, accessing to images that are recorded as evidence by investigators, police enquiries for a suspects photos from a central information data base, video streaming of actions at incidents being transmitted to law enforcement forces for taking action, forwarding suspects biometric data to a command centre, building blueprints accessed by fire fighters, firefighters equipped by personal monitors and location tracking devices, are all applications in the PPDR sector that need more bandwidth than that provided by current dedicated MCSs.

The increase in the data demands, as described above, has motivated the governmental and non-governmental organizations to look for alternatives to current solutions and options for delivering mission critical services over systems that are capable of providing the needed operational, functional, and technical requirements in the near future.

1.5.2 UK Emergency Services Mobile Communications

TETRA is used in UK for emergency services as part of Public-Private Partnership contract signed between The Home Office and Airwave Ltd. Based on this contract, the UK government pays around 400 million pounds every year for Airwave Ltd to use TETRA as a nationwide dedicated PPDR network providing PPDR personnel with a narrow band reliable service.

The UK government decided already to pursue a commercial replacement option for the current too expensive and limited emergency service communication system provided by Airwave. The Emergency Services Mobile Communications Program (ESMCP) aims to replace Airwave provided solution with a commercial broadband support service for PPDR based on commercial LTE. This migration from a reliable narrowband system to a developing broadband system introduces a lot of questions and debates in terms of feasibility of commercial systems and their ability of providing reliable solution working as a replica of current dedicated MCSs. Also, there are concerns regarding the Mobile Network Operators (MNOs) commitment to maintain the agreed Quality of Service for the PPDR use that should be mentioned in the Service Level Agreements (SLAs).

The decision was taken in UK to have transition in two years to a purely commercial LTE network that support both Mission critical and non-mission critical services over the Emergency Services Network. The research aims to have the new system compatible with the

current terminal equipment and applications, provide all the features that are supported by TETRA (DMO, PTT, group call, etc.), and at the same time provide a broadband and voice services at much lower cost. Hardening issues will be discussed in later sections of this thesis.

1.5.3 The Feasibility of Mobile communications as MCS

The mobile communication technologies has evolved significantly since early 1990's. 3GPP LTE release 11 is capable of providing broadband services to its users due to the advances in physical layer technologies such as Multiple Input Multiple Output (MIMO) Antennas, and Orthogonal Frequency Division Multiple Access (OFDMA). Moreover, LTE is optimized for minimum delay in both the Radio Access part and the Core Network part, which minimize the end-to-end delay of the service. However, although LTE can provide broadband services with low latency, it can only partially support the mission critical services requirements in the different sector applications. This is due to the many limitations and shortcomings of current mobile communication networks especially in the process of integrating end-to-end solution to support the mission critical operations will all needed requirements. The shortcomings and the need for hardening the overall system will be briefly introduced in the next subsection.

1.5.4 Commercial Mobile Technologies Concerns and Hardening need

The main shortcomings of current mobile communication technologies will be discussed more in later sections, for now it can be summarized as follows (Blom 2008):

- Voice Services Support: the lack of talk group communication, Push to Talk (PTT) and Direct Mode of Operation (DMO) support in nowadays mobile communications is considered a major issue for supporting Mission Critical Services over current mobile networks. LTE as broadband technology is considered data centric technology that is not optimized for voice services and need more effort toward supporting voice centric services, Voice over LTE (VoLTE) is contributing in this direction but needs a lot of efforts to integrate it as a solution in generic communication systems.

- Resilience: the limited resilience of the commercial mobile networks is unacceptable for mission critical applications in the three sectors. A lot of efforts need to be done to support the resilience in face of expectable events such power outage, natural disasters, or terrorist attacks. The resilience can be enriched by deploying backup and redundancy solutions in the access and core network parts.

- Pre-emptive capability: the PPDR applications need full access to the available resources whenever needed. Having a commercial system that is deployed for both conventional and professional use rises the resource allocation and scheduling challenge to give access to call of higher priority on expense of lower priority ones that need to be pre-empted based one certain rules.

- Coverage: the coverage of commercial mobile service is always enhanced for marketing purposes. The operators are interested in getting the maximum profit of their investment by building the network infrastructure and facilities near the densely populated areas with less coverage in rural areas. This would be a concern for mission

critical users who may need the service anywhere and anytime. Therefore, having a 100% coverage is of great importance for having nationwide MCS. Emergency Medical Services, for example, need ubiquitous coverage in their operations.

- Different Models: the need for different communication models is a need for having a reliable and resilient service. Air-Ground-Air (AGA) communication, for example, is of great importance for search and rescue operations in PPDR sector. Therefore, commercial mobile communication systems need to support AGA in addition to the Ground-Ground supported communication.

Based on the previously mentioned concerns, the need for having a hardened commercial system to work as a replica of current dedicated mission critical systems is of great importance for a reliable system with broadband services support.

The emerging of the LTE capabilities along with IP Multimedia Subsystems (IMS) based services such as Push over Cellular (PoC) requires further hardening and optimization efforts in order to meet the strict requirement of MCS. Hardening efforts for mobile technologies was investigated by the research community, In (TETRA MoU Association, 2004) the potential of General Packet Radio Services (GPRS) and Code Division Multiple Access (CDMA) for mission critical use was analysed by TETRA Association, the study shows that the latency of call set-up time is not acceptable for MCS use. In (Kanti, 2006) IMS was added on top of GPRS network to minimize the call-setup latency, which added advantages for group calling but with added call setup overhead. In (Balachandran, Budka, Chu, Doumi, Kang, 2006) and (Balachandran, et al. 2005) , IMS with PoC on top of UMTS and CDMA2000 feasibility was investigated and shows that the overall system latency was reduced but still not enough for critical communications. In (Blom, et al., 2008) and (IPWireless, 2012) the needs of the public safety community in HSPA and LTE was addressed.

1.6 PROBLEM DEFINITION

Based on what we discussed before, there are different deployment options for the mission critical systems, either to have a dedicated system or a commercial systems. Unfortunately, there are drawbacks for both options and many trade-offs that control the design and deployment process. What is important to know is that at the early deployments of dedicated mission critical systems, the voice and simple text messages were the only services that were used by PPDR for their critical missions, simply because mobile communication networks was at its early steps of evolutionary progress where the operators only thought of providing voice communication between users. Hence, when the dedicated mission critical communications was deployed, they also designed the whole system for voice and low data rate services which was fair enough at that time for providing the basic communication platform for the clients. However, due to the exponential increase of mission critical data use to support different new type of services and tasks, the need for broadband data along with the traditional basic voice service has become crucial for the mission critical communication systems.

Due to the increased demands for more bandwidth for the mission critical services, the need for access technology supported by a core network that is capable of providing broadband connectivity and scalable for high number of users along with broadband network support in the core is considered crucial requirement for any mission critical communication system.

1.7 SCOPE OF THE STUDY

End-to-end solution that incorporate all ISO/OSI model network model layer are of wide scope in all aspects due to the inclusion of many details in the design, modelling, evaluation and deployment stages. Technical standardization bodies, such as 3GPP, start a collaboration work with multiple institutions and research teams working separately and jointly to cover the needed layers development related to the standard. Application, Transport, Network, Datalink, and physical layer related standards are being modified and issued every year by the standardisation bodies. In our research, the proposed future MCS that is supposed to replace the current dedicated ones is like any other communication system that has multiple layers with potentiality of enhancing its performance at each layer or within two or more layers via cross layer optimization approaches.

This study will investigate the application and physical layer performance aspects and aims at proposing a cross layer optimisation approach between parts of features of the two layers. It worth to mention the layer, protocols, and technologies that will Not be covered to narrow the scope of the study. This thesis will not discuss the Transport layer performance and it is influence over the proposed solution (such as the variations between TCP and UDP usage), it is not supposed to study the data plane media streaming and its related application layer level protocols (media voice/video codecs, RTP, voice/video data QoS), the SIP message structure will not be altered but it is necessary to understand it for performance improvement purposes, operating systems software and kernel operation scheme will not be improved (such as OS kernel process scheduling), internal computer architecture design and performance issues will not be covered (such as multi core/multi-threaded core CPU performance impact), Security topics, such as IPsec protocol, will not be evaluated (although brief description of registration process authentication mechanism will be described), SIP server parser performance will not be enhanced, memory allocation and increasing CPU utilisation techniques will not be covered. In the physical layer, this study will cover part of the current 4G and future expected 5G physical layer techniques and methods, but it will not cover the antenna design, wave propagation, nor channel models aspects. Data link layer resource scheduling and multi user multiplexing evaluations and performance will not be covered, and finally network layer (IP layer) traffic routing, path discovery, and congestion issues will not be covered.

As shown in figure 1.1 the study will focus on the cross layer optimization of SIP signalling between the LTE-IMS interface in addition to IMS-Application Server interface. Optimization for less delay is needed to enhance the QoE for the mission critical services.

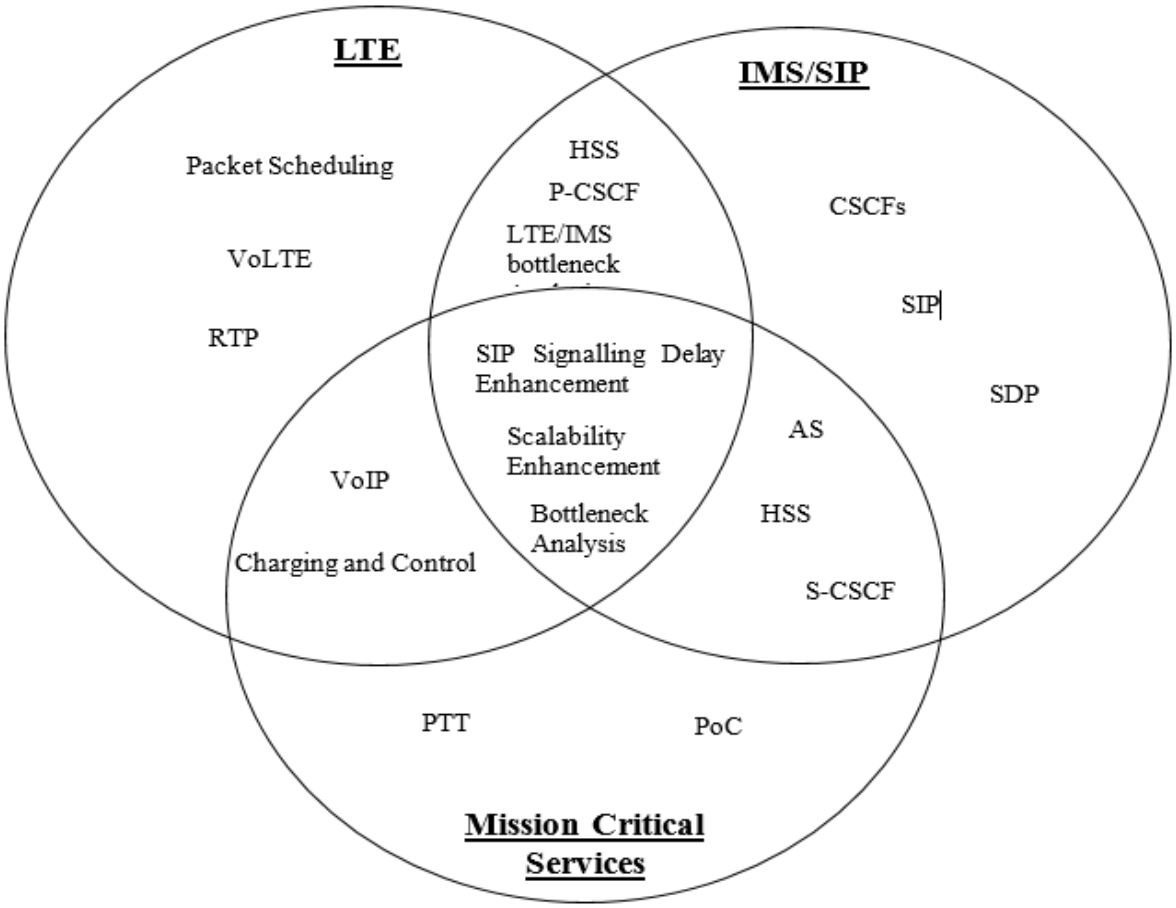


Figure 3.1 Scope of the Study

1.8 AIMS AND OBJECTIVES

The purpose of this research is to investigate the performance optimization issues for the mission critical communication systems that are deployed over the current LTE (4G), LTE-advanced, and 5G broadband communication Systems.

The emerging of the LTE capabilities along with IMS-based services such as Push over Cellular (PoC) based on Voice over LTE (VoLTE) system requires further integration and optimization efforts in order to meet the strict requirement of communication needed for emergency services.

In this research, the objective is to investigate more the SIP signalling over LTE to de-risk the compliance of any Mission Critical System for the End-to-End system performance targets. In addition to identifying the potential risks within the End-to-End system performance at the component performance level.

The research is contributing to the development of mission critical services over 4G and 5G broadband wireless networks with the deployment of a Large Scale Test program to de-risk the compliance of the Open Mission Critical System prototype product against the System and Software Design Description for the End-to-End system performance targets. In addition to

identifying the potential risks within the End-to-End system performance at the component performance level using related testbeds and simulation tools such as OPNET.

We can summarize the objective of the research as follows:

1. To have a clear understanding of MCS Technical functional and user requirements.
2. To identify and define a set of performance metrics that need to be measured for better understanding of system performance measuring criteria.
3. To propose a new framework model to be used in the proposed mission critical core network system.
4. To test the system over a testbed and collect statistics based on pre-defined performance measures.
5. To simulate the system over simulators and generate behavioural charts and statistics that reflect the performance of the default system.
6. To identify the areas of weaknesses in the system and suggest ways for performance improvement.
7. Validate the suggested and proposed improvement method over the testbed and simulation tool to get the new enhanced behavioural model of the system.
8. To compare between the two collected statistics before and after applying the proposed enhancement approaches for both the testbed and the simulation tool.

1.9 RESEARCH CONTRIBUTION

In this research, the added contribution to the field can be summarised in the following points:

1) Design a novel framework for IMS and its interfaces along with its interface connectivity with the access technology domain to enhance the signalling in overload scenarios in hope it will be able to increase the overall system scalability and reliability to be suitable for MCS communications and applications.
2) Define different modes of operation for the framework that shows the connectivity structure and its dynamicity against different overload or normal load conditions.
3) Analyse the interfaces inside IMS that connects different entities with each other and analyse the signalling traffic in different modes of operation and in different call scenarios to reflect all possible kinds of user end-to-end signalling options (Home caller/callee, visitor caller/call, mobile/fixed user, SIP INVITE/SIP non-INVITE traffic types, etc.). This include as well counting the DNS and lookup tables performance implications analysis.
4) Characterise the performance of the system in real time, The performance model then need to be simulated to check its validity for different or at least multiple connectivity scenarios under different traffic generation processes options. State flow simulations can be used to get a better idea of model performance.

5) Carry out a performance analysis study through general model performance metric evaluations that reflect the scalability and capacity in addition to the response time values.
6) The model can then be customised/adjusted based on traffic analysis of an experimental testbed collected results that reflect a special case real experimental system model and compared with the more general theoretically proposed model.
7) Add traffic monitoring and performance metric capabilities to IMS entities to be able to analyse the traffic in real time and extract the performance metrics of interest instantaneously in a log file.
8) Import the experimental results from the log file and control the operation of the IMS traffic routing inside the IMS based on the extracted real time results.
9) Compare the performance of both the experimental and modelled systems to decide possible improvement in the proposed framework operation.
10) Exploit the physical layer of LTE/LTE-advanced physical layer and present modified version of the initially proposed framework design.
11) Run physical layer simulations that shows performance challenges in the future 5G systems and decide the points of improvements via cross layer optimisation with the SIP application layer.
12) Present the cross layer optimization algorithm and show, via simulations, the possible performance improvement.

1.10 STRUCTURE OF THE THESIS

The thesis is structured in a way to let the reader understand the problem, get general background information about the topic, recognise the existing technologies, gain sight at the future vision and requirements, then have an idea about the detailed literature recent related literature studies, and finally get the proposed contributions and the evaluation methodology and results. In more details, the thesis is organised as follows:

Chapter 1: Introduction: an overview of wide literature coverage is presented in this chapter. The problem definition and scope of the study will identify and highlight in general the proposed contribution domain.

Chapter 2: MCS Protocols, Requirements, and Technology Background: In this chapter, the mission critical communication application and services requirements and evaluation of the compliance of current system and future proposed systems is analysed. The different access technology standards that has the potential to deploy MCS or already being used as MCS are described. Moreover, the SIP protocol operation and signalling details are described and its operation within IMS network is explained. The performance metrics related to both SIP and IMS performance are listed and described in detail.

Chapter 3: Literature Study and Related work: a detailed literature study of the standards and protocols that are related to the proposed MCS system structure are listed and investigated, then it is followed by classifying the literature into different classes based on possible performance improvement points. And finally the gap in the field and suggestion on how to fill the gap is listed and described.

Chapter 4: Methodology and Signalling Analysis

The systematic methodology approach to design the framework and evaluate the performance is explained in this chapter.

Chapter 5: Enhanced MCS Framework Design

The framework design details are explained in this chapter, systematic methodology was followed to describe the system detailed functionality and entities. Flowcharts and algorithms described precisely the signalling interfaces and execution order of the entire system.

Chapter 6: Performance Evaluation Results of the Framework Design

Simulations and experimental testbed using different scenarios was carried out in this chapter. The results were collected and discussed and analysed. New performance metrics were introduced to reflect the precise system performance. The evaluation mechanism was fully explained in this chapter.

Chapter 7: Conclusions and Future work

This chapter finishes the thesis with the conclusions made and potential future work extensions.

1.11 LIST OF PUBLICATIONS

A. Conference Papers:

- **Ashraf A. Ali**, Spyridon Vassilaras, Konstantinos Ntagkounakis, "A comparative study of bandwidth requirements of VoIP codecs over WiMAX access networks" NGMAST 2009 - 3rd International Conference on Next Generation Mobile Applications, Services and Technologies (2009).
- **Ashraf A. Ali,** Mazin Al Shamrani, Khalid Al-Begain "Evaluating SIP Signalling and QoS over LTE based Mission Critical Systems" NGMAST 15, Cambridge UK 2015.
- **Ashraf A. Ali**, Alhad Kuwedekar, Khalid Al-Begain "IP Multimedia Subsystem SIP Registration Signalling Evaluation for Mission Critical Communication Systems" 8th International al Conference on Internet of Things (iThings 2015) Sydney Australia 2015.

B. Edited Books:

- Khalid Al Begain, **Ashraf Ali** (Editors). "Multimedia Services and Applications in Mission Critical Communication Systems" May 2017, IGI-GLOBAL, USA.

C. Authored Chapters:

- **Ashraf A. Ali**, Khalid Al-Begain "Introduction to Mission Critical Systems and Its Requirements", Book Chapter (pages 1-18), Multimedia Services and Applications in Mission Critical Communication Systems, May 2017, IGI-GLOBAL, USA.
- **Ashraf A. Ali**, Khalid Al-Begain , IP Multimedia Subsystem and SIP Signalling Performance Metrics (pages 19-35) , Book Chapter, Multimedia Services and Applications in Mission Critical Communication Systems, May 2017, IGI-GLOBAL, USA.
- **Ashraf A. Ali**, Khalid Al-Begain, Session Initiation and IP Multimedia Subsystem Performance Evaluation (pages 36-49) Book Chapter, Multimedia Services and Applications in Mission Critical Communication Systems, May 2017, IGI-GLOBAL, USA.
- Mazin I. Alshamrani, **Ashraf A. Ali** , "Performance Metrics for SIP-Based VoIP Applications Over DMO" (pages 50-79) **,** Book Chapter, Multimedia Services and Applications in Mission Critical Communication Systems, May 2017, IGI-GLOBAL, USA.
- Mazin I. Alshamrani, **Ashraf A. Ali,** "QoS and Performance Evaluation for SIP-Based VoIP Over DMO" (pages 80-114)**,** Book Chapter, Multimedia Services and Applications in Mission Critical Communication Systems, May 2017, IGI-GLOBAL, USA.

D. Accepted Journal Articles Subject to Corrections:

- **Ashraf A. Ali**, Khalid Al-Begain, and Andrew Ware "Multimedia Services Key Performance Indicators Definition via SIP and IMS Registration Performance and Challenges Analysis", submitted to: The International Journal of Interactive Communication Systems and Technologies.
- **Ashraf A. Ali** , Khalid Al-Begain, and Andrew Ware "Scalability and Performance Analysis of SIP based Multimedia Services over Mission Critical Communication Systems " Submitted to : The International Journal of Interactive Communication Systems and Technologies

E. Accepted Conference papers:

- **Ashraf A. Ali**, Fatimah Al-Zahrani, Khalid Al-Begain, Andrew Ware "Performance Evaluation and Benchmarking of Mission Critical Medical Access Systems", IEEE HealthCom 2018. (Accepted on 19th of July 2018).

Chapter 2: MCS PROTOCOLS, REQUIREMENTS, AND TECHNOLOGY BACKGROUND

2.1 INTRODUCTION

In this chapter the main MCS related protocols are discussed, the protocol performance and signalling has a considerable effect over the overall system performance. The MCS requirements are discussed, the diversity of applications of MCS systems reflects over the requirements as well. The chapter will discuss the current technology implementation and how it meets the general set of requirements intended for an MCS.

2.2 MCS SERVICES REQUIREMENTS

The MCSs are designed to tolerate the consequences of natural disasters and abnormalities, and it is expected to be working to support the users on the ground, Otherwise it would be useless to use the whole communication system which is by nature design to operate in harsh environment. In this section, the general requirements of MCS will be discussed, then dedicated MCS implementations will be presented.

Services such as voice, data connectivity, messaging, push to talk, location and security services are the most important offered services provided by Public Safety (PS) Communication systems. There are set of challenges and requirements associated with each service. Voice services, for example, needs an acceptable packet loss ratio and can tolerate up to only 5% loss ratio for acceptable voice quality, moreover the end-to-end delay of voice streaming packets must be very small to make sure the two-way conversation is understandable. On the other hand, messaging services do not have the aforementioned delay concerns and the band width requirements are much less than other services such as voice and video. In this section, the requirements of different services over MCS will be discussed and explained.

2.2.1 Delay

There are many requirements for mission critical services over PSNs; one of the most important requirements is end-to-end time delay for sending the voice or data between system users. Delay is the most important requirement in PSN especially for real time services such as voice and video and non-real time services such as sending urgent push messages to report the occurrence of incident and to notify others within hundreds of mille-second time orders. There are different types of delay as shown in figure 2.1. For a voice call over the network, there are mainly two types of delays: the signalling delay such as the registration, call setup, and call termination delay. The data streaming delay is mainly the end-to-end delay and inter-packet delay or Jitter delay. It is crucial to minimize all the delays above to guarantee the best Quality of Service (QoS) and Quality of Experience (QoE) for the end users. In later chapters, we will investigate in depth the sources of delays for any service in both the data plane and control plane.

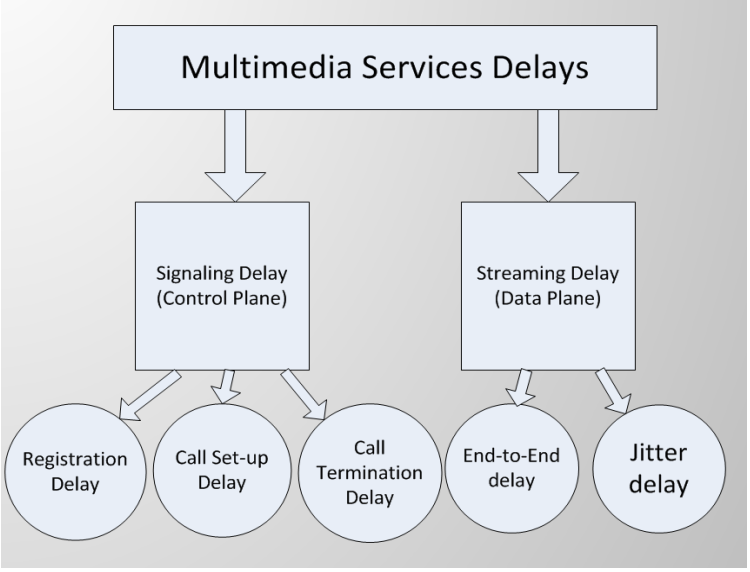


Figure 4.1 Multimedia Service Sources of Delay

2.2.2 Interoperability

Interoperability is needed for mission critical systems in places where different PPDR members from different organizations are all located in the same incident site, each has its own equipment's and they need not only to communicate within the same team only but also with other teams. For example, if there is a building of multiple stories set on fire, then it is expected to have on the ground of the site fire fighters to extinguish the fires, law enforcement members to investigate the trigger of the fire, and Emergency and Medical services members to help those who are injured on site. Clearly liaising between the different teams is needed to ensure interoperability for best utilization of available resources. Such interoperability can be achieved in general by following one of two approaches; the first approach is to use the same mission critical communication system by everyone. The second approach is to have different MC communication systems that are working under the umbrella of one common core standard that ensures providing the needed interfaces for seamless connectivity between the different systems. In (Balachandran et. al, 2006) a framework was proposed for a system that allows interoperability between multiple MCSs through a shared IP-core network, and interoperability with legacy mobile communication networks is presented in (Budka et.al, 2011). In addition to what is described before, there is also interoperability issues across borders between neighbouring countries that may use different frequency bands and MC communication systems.

2.2.3 Priority Handling

Moreover, priority handling is important for mission critical services especially if the infrastructure being used is shared with public users over commercial communication network where the need for employing pre-emption scheduling algorithms is crucial to guarantee the availability of resources for the PSN users at the appropriate time, clearly this means that less priority calls will be dropped from the queues if the system is not able to adapt all calls at the same time.

2.2.4 Resilience

Another important requirement is to insure the maximum resilience of the overall system, this implies that there should be backup systems in distributed environment to ensure reliability and to avoid the single point of failure. PSN should also be able to support Direct Mode of Operation (DMO) in which users can communicate with each other using their terminals in ad-hoc manner, this is important in scenarios where the access network may fail due to a catastrophic events that may be caused by terrorist attack or natural disaster. It is sometimes referred as all-time-available communication system which is, for Public Safety services, may be considered the tiny difference that may contribute in saving a precious life. In next chapter, we will see that the main challenge that makes the difference between the major types of public safety communications is the reliability of the whole MC communication system.

2.2.5 Security

Security is needed in any communication system; it is also of a great importance for any mission critical service where usually the messages being interchanged between PPDR teams is confidential and critical in nature. Hence, the need for a communication system that guarantees the end-to-end data integrity, confidentiality and resources availability against any possible attack is of crucial importance. Although deploying security framework for the mission critical communication system is out of the scope of this study, we will use it as a criteria for comparing different mission critical communication systems in later chapters.

2.3 MISSION CRITICAL SYSTEMS IMPLEMENTATIONS

2.3.1 TETRA

To ensure the strict requirements of the public safety communications, all governmental and non-governmental organizations nowadays rely on dedicated Public Safety (PS) communication systems due to their robustness, resilience, and reliability. Such systems ensure that, in most extreme and harsh scenarios, the network will be able to offer the service within the minimum acceptable service requirements.

As mentioned before MCSs can be mainly classified as dedicated or commercial systems. Having a dedicated PS communication system means that the system was built specifically to meet certain needs, and it also means that the operational and capital expenditures of deploying the PS system would be much higher than any other general purpose commercial communication system even though the license fees of the spectrum is usually waived out for the sake of public benefit. In this section, two dedicated MCSs and one of the expected commercial MCSs will be presented.

The Terrestrial Trunked Radio (TETRA) (Terrestrial Trunked Radio, 2003), or what is used to be named as Trans European Trunked RAdio, is a set of standards developed by the European Telecommunications Standardization Institute (ETSI). TETRA is a standard for a private Mobile Digital Radio System to meet the needs of Public Mobile Radio (PMR) organizations. TETRA is a communication system used by professional governmental and non-governmental agencies and operate independently from other commercial networks. It mainly uses Time Division Multiple Access (TDMA) with four channels each of 25 KHz bandwidth to support

both voice and data communication in point-to-point or point-to-multipoint manners. Figure 2.2 shows the general architecture of TETRA communications system. In addition to the device to base station radio interface, the handheld devices are able to communicate directly with each other without the need of the base station or core infrastructure. This device to device communication is known as Direct Mode Operation (DMO) and it is mostly needed in scenarios where public safety members are located in the same site, and there are no available nearby base station or the base station is unavailable due to an attack or natural disaster. DMO enhances the reliability of TETRA systems and widely used in nowadays deployments. On the other side, devices may access the core network using the base station in what is called Trunked Mode Operation (TMO) which uses the Switching and Management Infrastructure (SwMI) that is made of multiple base stations.

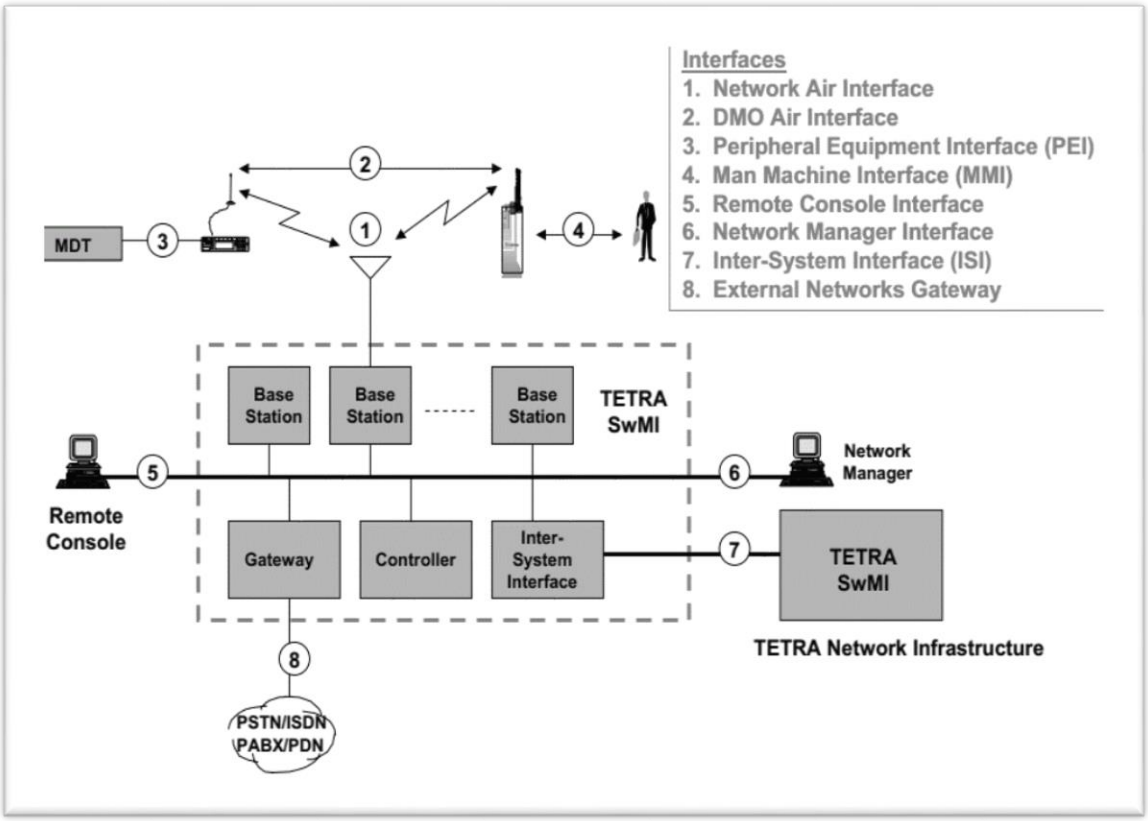


Figure 2.2 TETRA System Architecture (European Telecommunication Standardization Institution, 2013)

2.3.2 MCS Requirements compliance over TETRA

As mentioned before, TETRA (Terrestrial Trunked Radio, 2003) was designed and built to be a critical communication system for public safety duties and nowadays it is widely deployed in many countries. Hence, it meets all the requirements for MCS, TETRA has short call-setup time ,as shown in figure 2.3, where 85% of the calls need less than 500 ms (Terrestrial Trunked Radio, 2003) that is acceptable for real time services. It also provides bidirectional authentication between terminals and the core infrastructure in addition to air interface as well

as end-to-end encryption to meet the security requirements. It also supports DMO and TMO operations that enable point to point and point to multi point communication in both modes. This supports the reliability and resilience for MCS. TETRA has a wide coverage due to the low frequency of the carriers, and this will increase the range of maximum distance for communication either in DMO or TMO. Devices can work as a relay or repeaters to provide access for other devices to access the network in ad-hoc manner in scenarios where some terminals may fall outside the communication range, this feature enhances the both the availability and accessibility requirements of the MCS. TETRA provide transportable network solutions that are carried over vehicle that supports both reliability and capacity requirements. It also supports interoperability with other communication networks through gateways for wider accessibility domains and better interoperability enhancement between different domains and working groups. Finally, there are different levels of priorities that are defined in TETRA, pre-emptive call scheduling based on the priority class will guarantee that the more critical calls pass through the network with less priority calls schedules automatically at the bottom of the queue to be served afterwards.

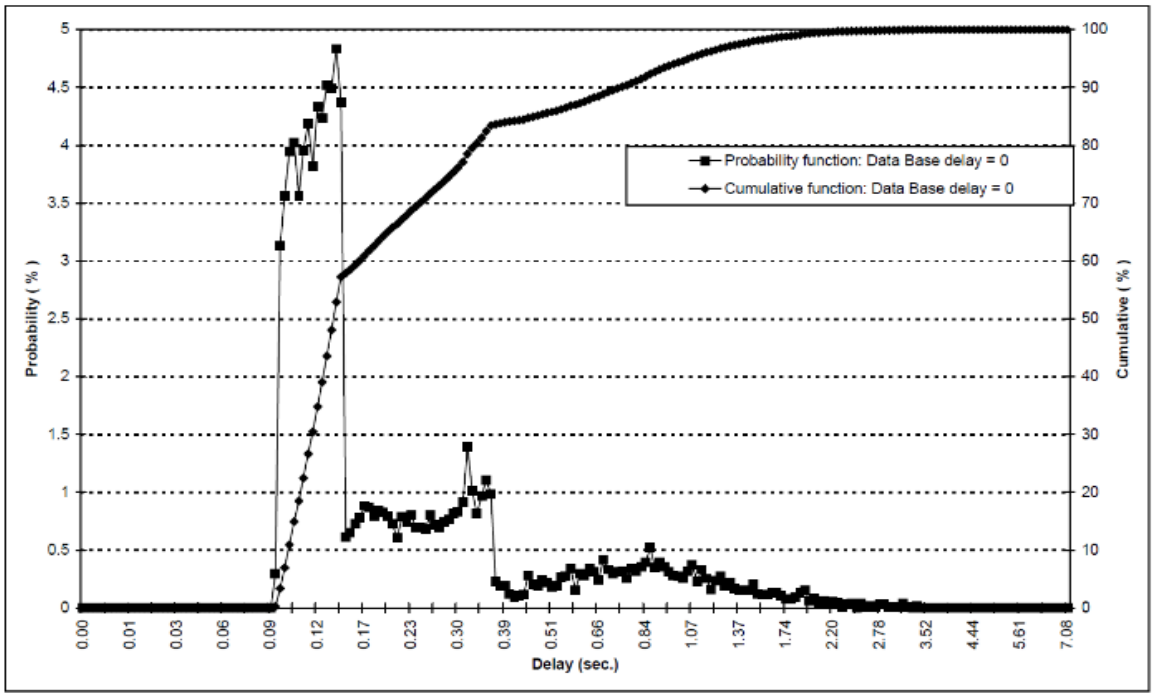


Figure 2.3 Call Set-up Delay PDF and CDF (Terrestrial Trunked Radio, 2003)

TETRA Release 1

TETRA Release 1 or TETRA Voice plus data (TETRA V+D) is the original standard developed by ETSI; it forms the basis for all other TETRA systems. Release 1 of the standard introduces the voice plus data (V+D) service for TETRA systems, the Direct Mode of Operation, and the Packet Data Optimized (PDO).

TETRA uses Time Division Multiple Access (TDMA) for user multiplexing. Figure 2.4 shows the TDMA structure and the types of frames starting from the hyper frame level down till the

subslot unit. It is important to note that the bandwidth utilization in both time and frequency domains is necessary to maximize the symbol rate or bit rate available for each user. In Release 1, π/4-shifted Differential Quaternary Phase Shift Keying (π/4-DQPSK) is the only modulation scheme used. Hence, the maximum bitrate available for users can be calculated knowing that the symbol word depth for (π/4-DQPSK) will be 2-bits per symbol and the user slot time is 14.167 ms over the 25 KHz bandwidth channel with a data rate of 36 Kbps.

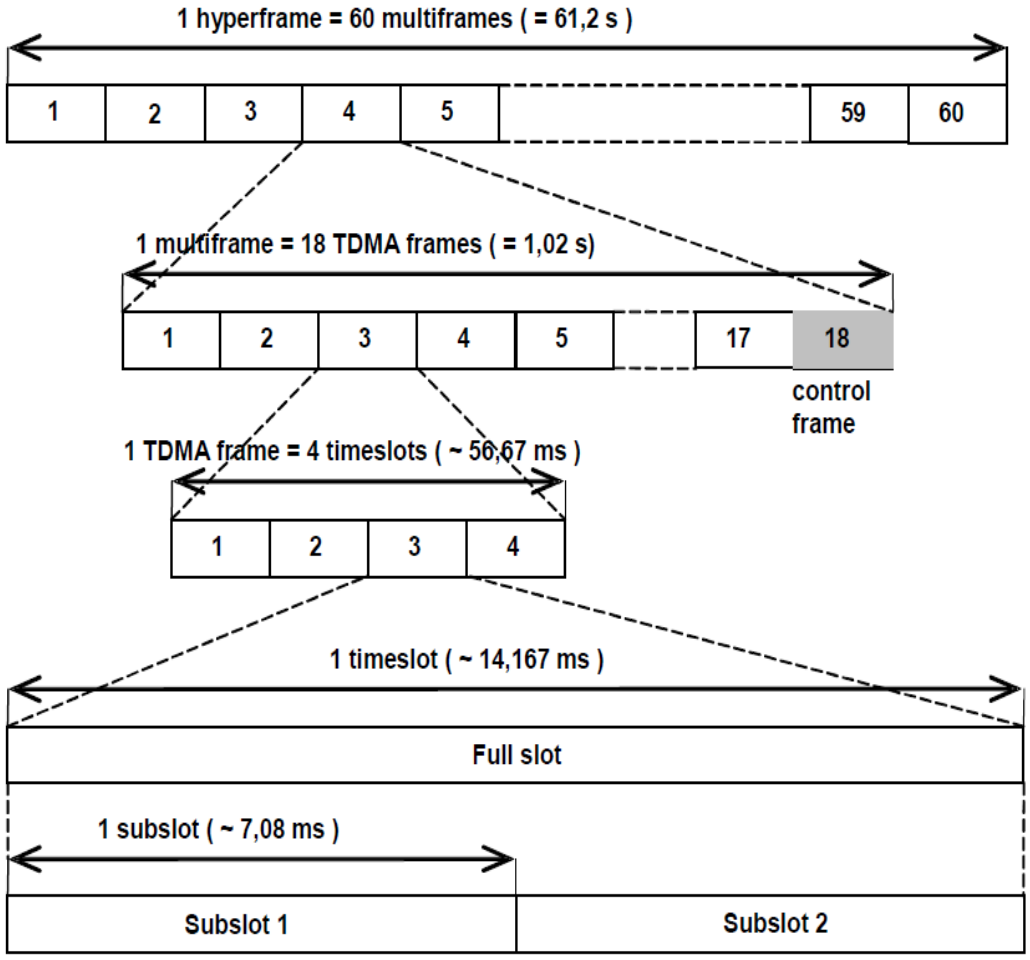


Figure 2.4 TETRA Frame Structure (Terrestrial Trunked Radio, 2003)

TETRA Release 2

I this release, the access using TDMA is done via four physical channels per carrier and there are multiple modulation schemes that is pre-allocated based on the channel width. For Phase Modulation (PM) the carrier bandwidth is 25 KHz, whereas for Quadrature Amplitude Modulations (QAM) the carrier bandwidths are 25 KHz, 50 KHz, 100 KHz or 150 KHz depending on the scheme of QAM modulation. Table 2.1 shows the symbols per slot or sub-slot that are available per user or physical channel at certain instant of time. The symbol rate using PM is $(18 * 10^3)$ Symbols/s and for QAM it would be $(1.2 * 10^3)$ Symbols/s.

Table 2.1 Symbols Capacity in TETRA

	Number of symbols	
	Phase modulation	QAM
Slot	255	34
Subslot	127,5	17

Now we can calculate the bit rate based on the symbol rate knowing the modulation scheme used in both the PM and QAM. In PM there are two modulation schemes; π/4-shifted Differential Quaternary Phase Shift Keying (π/4-DQPSK) or π/8-shifted Differential 8 PSK (π/8-D8PSK). Hence, the symbol word depth for (π/4-DQPSK) will be 2-bits per symbol and the overall data rate is 36 Kbps, similarly the symbol word depth for (π/8-D8PSK) is 3-bits per symbol and hence the overall data rate is 54 Kbps.

For QAM modulations, there are three modulation schemes 4-QAM, 16-QAM, or 64-QAM, and there are 8 sub-carriers per 25 KHz being used. Each subcarrier has a symbol rate of 2400 Symbols/s. Hence, the maximum available bandwidth for users is shown in Table 2.2.

Table 2.2 Gross bit rates for QAM Carriers (Kbit/s)

Modulation type	Carrier bandwidth			
	25 kHz	50 kHz	100 kHz	150 kHz
4-QAM	38,4	76,8	153,6	230,4
16-QAM	76,8	153,6	307,2	460,8
64-QAM	115,2	230,4	460,8	691,2

2.3.3 APCO P25

Similar to the role of TETRA, the Association of Public-Safety Communications Officials-international (APCO) started what is called Project 25 (P25) (Project 25, 2016) which is a standard for public safety agencies in North America. P25 was designed to address the needs for a professional Digital Mobile Radio system used by public safety agencies to meet all the mission critical service requirements. It supports DMO and the conventional, trunked mode of operations.

APCO evolved in two phases, in phase 1, the FDMA was used over channels of 12.5 KHz bandwidth that provides a maximum of 9.6 Kbps per user of which 4.4 Kbps for voice data, 2.8 Kbps for Forward Error Correction (FEC) and 2. Kbps is for signalling and control functions. On the other Hand, Phase 2 was developed for better spectrum utilization. It uses TDMA with two slots for medium access. A more efficient voice codec that require 6 Kbps for voice data, error correction and signalling.

In this study, P25 will not be investigated due to the similarity with TETRA and also because it has the same performance aspects of TETRA that need to be compared with the more generic proposed solution that will be presented in the next subsections.

2.3.4 Long Term Evolution (LTE)

2.3.4.1 Introduction to LTE

The Long Term Evolution (LTE) standard was developed by the Third Generation Partnership Project (3GPP) (3rd Generation Partnership Project, 2008), it started in Release 8 and then Release 9 till Release 10 for LTE-Advanced. It is considered a normal evolutional step of the mobile technologies but using revolutional communication techniques at the physical layer to allow higher bandwidth and less latency for the end users. It emerged due to the fact that users nowadays are more mobile, need more bandwidth, and their running services need less latency for better interactive and hassle free performance. Moreover, LTE was designed for end-to-end all-IP-Connectivity between end users for simpler design architecture and to avoid the multi-domain conversion added overhead.

There are two parts of LTE Communication Network as shown in figure 2.5; the first one is the Radio Access Network (RAN) part which is known as Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) that has enhanced performance compared with Universal Mobile Telecommunications System (UMTS) that was used in the 3G communication networks. Evolved UMTS (E-UMTS) is mainly responsible for managing the whole radio stack signalling between the access point that is called evolved Node B (eNB) and User Equipment (UE). The other part is the core All-IP-Network part which is called the Evolved Packet Core (EPC) that is mainly responsible for managing the bearer services, mobility management, and interconnecting the interfaces and gateways with other domains and entities. Fig. 1 shows the general architecture for LTE communication network

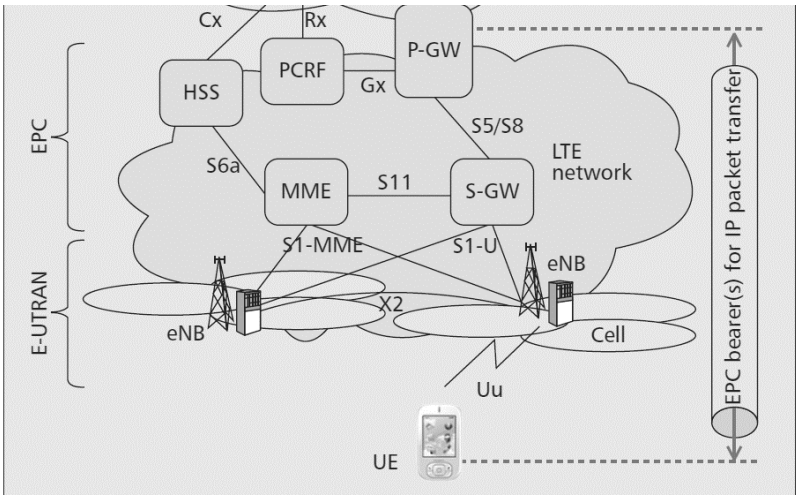


Figure 2.5 General LTE Architecture (3rd Generation Partnership Project, 2008)

2.3.4.2 LTE as a Mission Critical System

LTE is supposed to be the first interaction point with the users as shown I figure 2.6, it is a strong candidate to be considered for the Mission Critical and Public Safety Communication Systems due to its ability to provide the end users with the broadband services capacity and other MCS requirements (Doumi et. al., 2013). Delay and bandwidth are very important because nowadays services types need more bandwidth and less latency for sufficient QoS and reliable service. There are many challenges that face the LTE as MCS deployment such as supporting DMO and providing the needed interfaces to insure interoperability with already existing MCSs, in addition to supporting group voice calls over IP packet switched structure. But at the same time the All-IP-Network flat architecture is considered a big advantage for LTE to overcome the interoperability and complexity issues that may emerge in the proposed project. Moreover, the spectral flexibility and the higher spectral efficiency allows for more bandwidth utilization and better resource allocation for the end users which enhances the QoS provided, (Simic, 2012) highlights LTE capabilities in terms of providing different QoS classes at different levels.

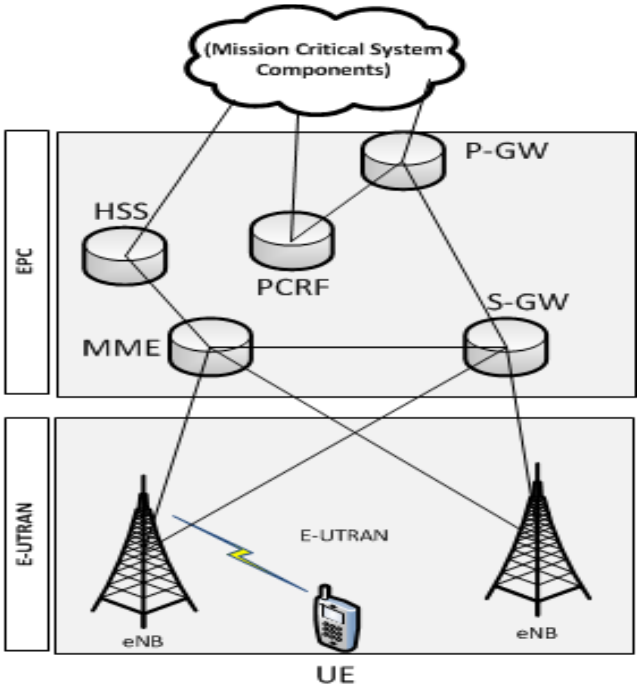


Figure 2.6 LTE as MCS

LTE provides a set of service preferences for the user to meet a certain level of service requirements, and the requirements affect the scheduling and queueing user data, priority and pre-emption capabilities, and the access control treatment. All the requirements above are out of the scope of the bearer domain and managed by the access technology. At the access level, there are 16 priority classes that may be dedicated to public safety users to overpass the network overload access issues and at the network level, the Evolved Packet System (EPS) bearer, which is a logical channel between the UE and the P-GW at the far edge of the EPC, has two types; the first one is the Guaranteed Bit Rate (GBR) bearer where the user has a reserved resource during admission, and the Non-Guaranteed Bit Rate (None-GBR) bearer that is based on best effort service without guarantees. There is also the Allocation ad Retention Priority

(ARP) which determines the priority class of the connected bearer in addition to two other flags; the pre-emption capability flag, that determine if it is allowed to pre-empt another lower priority bearer. The pre-emption vulnerability flag, which determines if the bearer may be pre-empted by another higher priority bearer. The APR will facilitate the decision of managing the bearer connection in overload conditions. In order to run MCS, the LTE needs to be coupled with the IP Multimedia Subsystem (IMS) to create an environment capable of supporting voice and video traffic in a shared packet data network. The next section speaks about IMS and its coupling with LTE.

2.3.5 5G Technologies for MCS

2.3.5.1 Introduction

The need for a less end-to-end latency and a more reliable communication system is crucial for future mission critical applications and systems. The future fifth generation communication systems is expected to adapt more real time applications that need higher data rate and less latency with more reliable structure compared to the current communications system in the market. Although low latency is needed in nowadays system to ensure having a usable applications that are sensitive to packet delay variations (jitter delay) and end-to-end packet delay, still the need for even less delay constraints is needed due to the introduction of Human to Machine (H2M) and Machine to Machine (M2M) which require a virtual zero latency communication systems. 5G communications systems is expected to provide the platform for such newly introduced applications.

2.3.5.2 MCS Future vision

5G use cases in different sectors is strongly related to its latency and reliability requirements. In certain applications, such as the remote surgery robotics, the need for less than 1 ms end-to-end latency and a very reliable communication system of Block Error Rate less than 10 ^-9 is essential due to the critical nature of the task assigned to it. On the other hand, augmented and virtual reality applications has less tough requirements by having End-to-end latency less than 5 ms to avoid cyber sickness symptoms and reliable enough systems to detect failures.

The diversity of applications in the newly developed applications will introduce a new set of requirements that are not yet satisfied in current 4G communication systems. 5G is not yet fully standardized as the definition of services and requirements is not yet clear, it is easy to provide a definition that embed optimum performance and perfect communication system, but at the same time it is not easy to hardcode in a standard using the best of currently available hardware and software technologies in the market. Everyone is looking for an “Extreme Mobile Broadband” in which what is referred as “unlimited experience” that will only be satisfied with 10 Gbps peak data rate and 100 Mbps as a guaranteed bit rate (GBR) taking into account the traffic will be 10000 more compared to current 4G technologies due to the introduction Internet of Things (IoT) and Internet of Everything (IoE) applications which significantly increase the need for a more scalable networks.

Critical Machine Communications in the future 5G communication systems needs instant action from both end user and backbone network to ensure that the task is finished within a

very strict time limits that cannot accept delays similar to VoLTE demands for example. Such systems need what is referred as "Ultra reliable" communications with less than 1 ms radio latency. As mention before, the introduction of IoT and IoE has introduced a new traffic that is generated from low cost devices. The user may have 10 to 100 more devices that are distributed in home and work environment and designed to signal and connect with the access technology seamlessly and without human interaction, some of such devices can be connected on batteries that have a long life span up to 10 years. Such changes in human interaction and life experience has introduced the concept of "Massive Machine Communication" to the future 5G communication systems.

Table 2.3 summarizes the expected services along with use case examples and general requirements. To meet the requirements mentioned before, the need to improve the radio access and core network functionality in addition to having a redundant infrastructure less is crucial for future 5G technologies. Ultra reliability, for example, can be achieved by deploying massive MIMO concepts (Mukherjee and Beard 2017). Device to Device communication (D2D) or Direct Mode of Operation (DMO) will also contribute toward a more reliable communication systems goal by having a redundant communication path between devices in disaster scenarios where the infrastructure of the communication system may collapse. It also increase the coverage, availability and scalability of the system.

Table 2.3 5G Applications Requirements

Application	**Scalability**	**Data Rate**	**Delay**
Extreme Mobile Broadband	Thousands of users per cell	>10 Gbps peak rate	<5 ms
Massive Machine Communication	Thousands of low cost devices	>1Gbps based on the application used	<3 ms for real time response
Critical machine Communication	Hundreds of machines in a site	Low Data Rate for control Signalling	<1 ms radio latency

2.3.5.3 MCS Applications of special requirements in 5G systems

Autonomous Vehicles

Self-drive cars is being a hot research topic for many researchers and companies. Saving lives, changing human's lifestyle, insurance companies' investments, improving road utilization, reducing journey time, and avoiding road congestion are among many factors that are pushing toward fully automated driven vehicle experience in the near future. It is expected that by 2025 75% of vehicles on motor ways will be autonomous. The communication system need to be a

very reliable system to avoid life threating scenarios, it need to be of end-to-end latency that is less than 10 ms to ensure reacting a timely manner for humans safety during the journey.

Augmented/Virtual Reality

In Augmented Reality, the real view of the user's eye perspective is enhanced with images and real data which could be enriched with multimedia services that are displayed based on user position or vision. On the other hand, in Virtual Reality, the user is experiencing non real environment that involves multimedia content in forms of images, background music, and videos in interactive manner to synthesis real life or even imaginary scenarios. In both cases, Augmented Reality and Virtual Reality, the user must enjoy the experience instantly without delays, which implies that the network and access technology must be capable of providing end-to-end delay of less than 5 ms to avoid any nauseating or uncomfortable user experience. In addition to the delay limitations, the need for high data rates to be able to mix all the multimedia contents for one user perspective to best emulate real life experience is needed, such high data rates need a communication system able to adapt many users who are expected to use Augmented and Virtual Reality applications as part of their everyday life routine.

The need for a reliable system is clear in scenarios in which PPDR forces may use Augmented and Virtual Reality to accomplish their tasks more efficiently and accurately. Firefighters for example, may use the Augmented Reality to reflect the blue print of the building over the real position of the team on ground to be able to recognize the safety zones and best ways to enter and exit the building, they may also get ambient temperature map of the whole building in real time to better estimate the risks and evacuation path need to be followed. Police enforcement forces may use the Augmented Reality to identify the suspect using face recognition and detection techniques in real time before arresting process occur.

Remote Robotics

In Remote Robotics, the machines are supposed to work remotely to replicate human expertise in certain fields. Surgery Robots for example, are designed to work remotely and controlled by MD doctor in scenarios where it is either difficult to be in surgery site or it is dangerous and risky for doctors lives to be there. Robotics are sent by PPDR in dangerous situations or by firefighters to get into high temperature zones that humans cannot tolerate. Such applications require a highly reliable system with BLER up to 10 ^-9. Similarly, such systems need to be responsive enough to replicate the human interactions, delays should not exceed 1 ms to enjoy the haptic experience and interaction.

Industry Control and Automation

Robotics in industry are used to accomplish the tasks the needs accuracy and to be completed in a timely manner, it needs a reliable communication system due to the critical nature of its duties, it needs a low end-to-end latency of less than 1 ms to keep the production loop synchronized . Most production flows nowadays relies on wired technologies to ensure the delay and reliability requirements. However, the future 5G technology, is expected to replace the wired structure to eliminate the cost and simplify the production flow setup and adds flexibility to change the production flow whenever needed.

2.3.5.4 Proposed Enhancements in 5G Communication Systems

In order to satisfy the stringent applications requirements introduced in the previous section, a revolutionary enhancements in current 4G systems need to be deployed in order to be counted as part of the next 5G technology solutions. In this section, Enhancements proposed in (Paper 2016) in Radio access, Frame structure, and core network will be presented.

Radio Access

The radio access technology is directly connected to the end user, it has a great impact on the overall and end-to-end communication system performance. End-to-end latency and overall system reliability and resilience is affected by the access technology specifications. Signal to noise ratio (SINR) is a measure of the channel quality. All User Elements (UE) report back periodically to the base station, or what is referred as evolved Node B (eNB) in 4G systems, the SINR to let eNB decide the best channel to use for the downlink traffic. To increase the reliability of the overall system, the system need to pick the channels with highest SINR and therefore least Bit Error Rate (BER) is crucial to enhance the overall system reliability. Microscopic diversity and Interference management are among the proposed techniques to enhance the reliability by altering the SINR.

Diversity

As described before, due to the SNR and BER dilemma, the need to get use of every power unit to enhance the reception quality and improve the BER measures is considered vital for a cost effective, less power hungry and efficient solution. Multi Input Multi Output technology (MIMO), which was first proposed in IEEE 802.16 (WiMAX) technology, is the key technology used to enhance the diversity between the transmitting and receiving unit. 2X2, 4X4, 8X8 and even much higher schemes which are referred as massive MIMOs are used in different technologies nowadays. 2X2 is the most common scheme used in nowadays technologies such as LTE. The 8X8 scheme was proposed in the latest 3GPP release with a full dimensional MIMO schemes.

When referring to “diversity”, it may be referred as combination of different technologies with redundant entities to provide a more reliable end-to-end connectivity. Apart from the physical layer transmission diversity (MIMO), there are other ways for providing the multi-connectivity principle in the higher layers for a lower latency and higher reliability. Combining multiple radio access technologies, such as the combination of LTE and 5G along with multiple cells and nodes with coordinated transmitters and receivers, is considered one of the proposed solutions for the future MCS. the current 4G communication systems that rely on LTE technology has a Break-Before-Make interruption time of 55 ms, which is the time needed to re-establish the connection and allocate the needed resources between end users after being interrupted unexpectedly or during the vertical or horizontal handover process . For the 5G systems, a virtual “Zero Break-Before-Make” time is needed to satisfy the “Ultra Reliable” system constraints.

Interference Management

Interference management is considered as another way to improve the SNR. Similar to the noise, the neighbouring interference from the base station or terminals reduces the quality of the signal and therefore degrades the SNR.

2.4 SIP AND IMS PERFORMANCE METRICS

2.4.1 Introduction

In this section, the most important protocol performance issues that control the operation of end-to-end systems and have implications on its overall performance are presented in this paper. There are mainly two types of protocols: control plane protocols and data plane protocols. The Session Initiation Protocol (SIP) is one of the control plane protocols and was fully standardized by the Internet Engineering Task Force (IETF) in RFC 2543 for the first version of SIP 1.0 and in RFC 3261 for the second version of SIP 2.0 (Rosenberg 2002). The SIP is a communication protocol that operates over IP and is used for the signalling of real-time multimedia services such as voice, video and non-real-time services (for example, text messages and presence notifications). The protocol, which is text-based, mainly defines the signalling order between end users for call initiation, termination, in addition to instantly modifying the call setup as needed. It is also used for registering the users before the call is initiated.

In this thesis, the SIP message headers and signalling details will not be presented. But the performance issues and the challenge of enhancing SIP services performance will be highlighted and briefly discussed. Figure 2.7 shows the signalling diagram for registering users, initiation and termination of a call between caller, callee, and back to back SIP server.

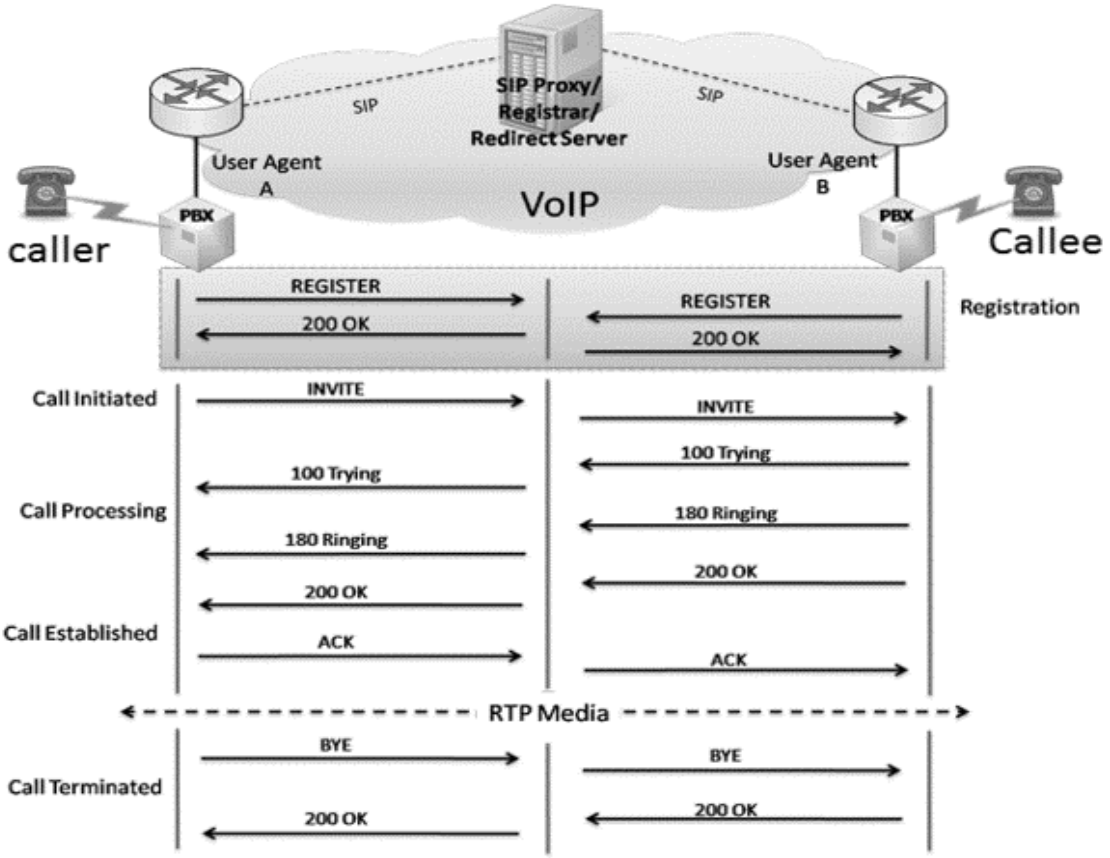


Figure 2.7 SIP Signalling Diagram

2.4.2 SIP Performance issues

The protocol related performance metrics need to be identified to determine the way SIP is utilizing the system resources and how to maximize it. Moreover, the architectural design challenges need to be targeted to enhance the SIP performance. Some of the protocol-related metrics in addition to implementation related metrics is discussed in (On SIP Performance, 2004), it shows different set of tests that measures the processing time for SIP messages, memory allocation, thread performance, and call-setup time. The results show that the performance of the proxy server changes by varying SIP related parameters and thus affecting the number of calls by seconds that the proxy server can handle at a time. It also shows that the performance of SIP related architectures, such as the IP Multimedia Subsystems (IMS) that will be presented later, is more affected due to the heavy dependence of SIP signalling and SIP messages structure compared with a simple Proxy/Registrar Server. Furthermore, it is important to note that the performance of SIP signalling is significantly affected by the delay at different stages of registration, call initiation, and call termination processes. The performance of SIP signalling will also affect the QoS of the offered service. Hence, the need to define the metrics that identify the performance measure for SIP is crucial for evaluation and performance comparison purposes.

IETF proposed the criteria for the end-to-end SIP performance measures in RFC 6076 (Malas and Morton, 2011).

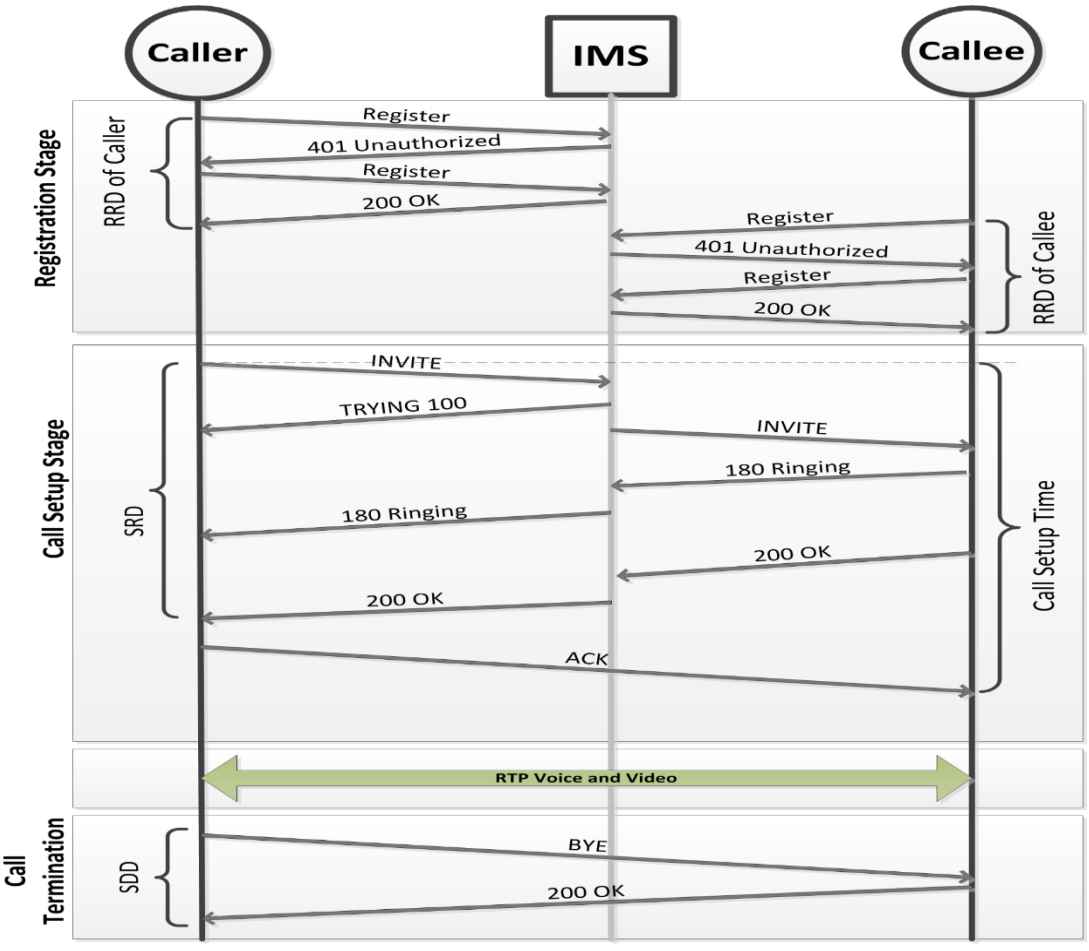


Figure 2-8 SIP signalling flow and performance metrics.

Due to the lack of an SIP benchmarking standard to define the baseline performance of SIP signalling, RFC 6076 was developed to articulate the performance metrics for SIP in VoIP applications in order to provide Key Performance Indicators and Service Level Agreement (SLA) indicators for best network resource utilization and best end user Quality of Experience (QoE). As shown in figure 2.8, the main metrics defined in RFC 6076 are the Registration Request Delay (RRD), Session Request Delay (SRD), and Session Disconnect Delay (SDD). RRD is the time needed for the user to finish the registration process successfully. SRD is the time needed to get a reply from the server side regarding the requested call setup from the user side; it is counted for both successful and unsuccessful call requests. If the call requested was successfully set then the call setup time will simply be the SSD plus the acknowledgment sending time. The SDD is the time difference between sending the BYE message from the user side and the time of receiving 200OK confirmation from the server.

In this research, the call setup delay is used to measure the system performance due to its importance in real-time multimedia services in general and in mission-critical communications specifically. The QoE for SIP-based systems is affected significantly by the call setup value. The mean time of call setup values can reach up to 800 ms (ITU-T 2004). However, for LTE-based mission critical systems (designed to work as a replica for traditional dedicated mission critical systems such as TETRA), the average call setup time needs to be within 500 ms delay.

Figure 2.9 shows the average access time of different technologies. The call setup time is normally a multiple of the access time due to the enforced handshaking process.

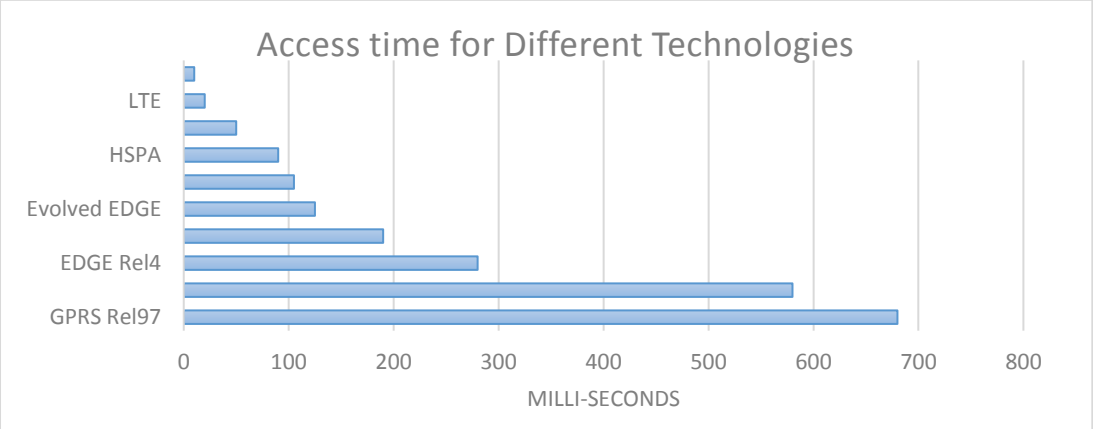


Figure 2.9 Access Time Evolution for Different Technologies

2.5 IP MULTIMEDIA SUBSYSTEM (IMS)

Legacy mobile communication networks and the next generation systems are able to provide end users with new sets of applications and services. There is a need for a system that sits on top of the access technology domain to provide the required signalling for multimedia services. IP Multimedia Subsystems (IMS) form the core part of Next Generation Networks (NGNs). They play an important role in helping the service operators merge the multimedia services in cellular networks and to provide end users with key features and services within Quality of

Service (QoS) levels set by operators. These multimedia services such as messaging, instant voice, video conferencing, group management, and push services, all rely on IMS to control the signalling and the state of the service before initiation, during the service and after service termination.

IMS was designed and standardized by 3GPP to provide multimedia services over mobile communication technologies beyond GSM (3GPP 2005). The IMS is used for delivering the IP multimedia services between users and service providers; it gains its importance as an architectural framework for multimedia communications. The SIP is the principal signalling protocol used within an IMS to create, modify and destroy multimedia sessions. As defined by the standard, an IMS functions as an interface between the service/application layer and the transport layer, enabling the service providers and operator to control the user QoS based on its subscription profile. Moreover, it works as a hub point for the entire SIP signalling that needs to take place before, during, and after the call. To ensure overall integrity, the different control functions are connected through a set of interfaces. Figure 2.10 shows IMS architecture.

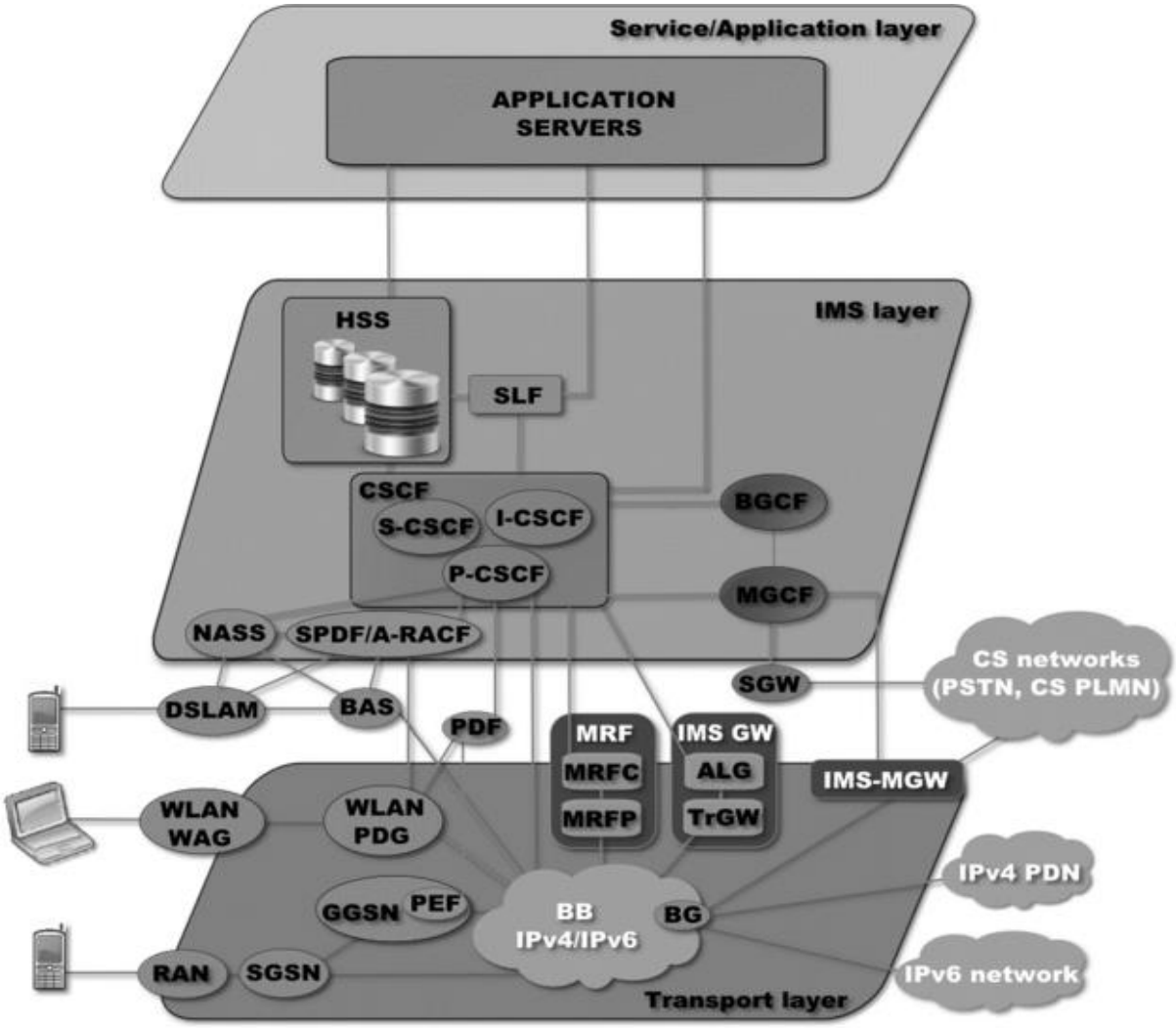


Figure 2.10 IMS Architecture

Users' subscription-related information is stored in the Home Subscriber Server (HSS) that performs authentication and authorization functions for users. The HSS is also responsible for

updating user registration status records. The Call Session Control Functions (CSCF) are responsible for handling the SIP signalling messages and packets in the IMS. The Proxy CSCF (P-CSCF) is the entry point into the IMS system and all SIP messages flow through the P-CSCF. The P-CSCF may also apply security or compression algorithms to the received traffic in addition to providing quality of service control and bandwidth management. The interrogating CSCF (I-CSCF) is one of the main elements of the IMS systems. It is used during the registration process when the UE does not know which Serving CSCF (S-CSCF) should receive the request. The I-CSCF interrogates the HSS to obtain the address of the appropriate S-CSCF that should process the request.

The S-CSCF performs session control functions that interface with the HSS to check and download the user profile information. It assigns the Application Server (AS) for the user for further services and enforces the operator policy control. The SIP AS has an SIP interface with the S-CSCF and is used to host specific IMS services. After the registration process is completed and the S-CSCF is allocated to the UE, the I-CSCF is no longer used for any further communication. All future communication happens between the UE, the P-CSCF, and the S-CSCF.

In the next sections, the performance of IMS, which is critical in affecting the overall system performance due to its hierarchal position in the stack and core functional role, is discussed.

2.5.1 IMS Performance Issues

After migrating from the circuit switched Second Generation Mobile Networks (2G) toward the packet switched domain of fourth generation (4G) communication networks and beyond, the need for supporting multimedia services in all IP network infrastructure is essential to ensure the convergence of data, multimedia services, and mobile networks technologies. IP Multimedia Subsystems (IMS) developed by the Third Generation Partnership Project are a key part of Next Generation Networks (NGN) that are responsible for providing and controlling the multimedia services in the packet switched domains. As defined by the standard, the IMS operates as an interface between the service/application layer and the transport layer that enables the service providers and operator to control the user QoS based on its subscription profile. Moreover, it works as a hub point for the entire SIP signalling that needs to take place before, during, and after the call. For this purpose, there are different functions that are connected by interfaces to ensure operational integrity.

In the protocol stack, the IMS resides between the application layer and the transport layer (see figure 2.11). Therefore, the performance of the IMS affect mission critical applications that have strict requirements due to their special operating nature and associated tasks. End-end delay of the access technology components along with the IMS entities affect the end-to-end signalling and QoE offered to the end user. Having a generic system that has IMS core network serving both legacy mobile users and mission critical system users will introduce overload, especially at the IMS side. In this thesis, evaluation of IMS and SIP performance metrics and KPIs was investigated scenarios with increasing number of clients was investigated

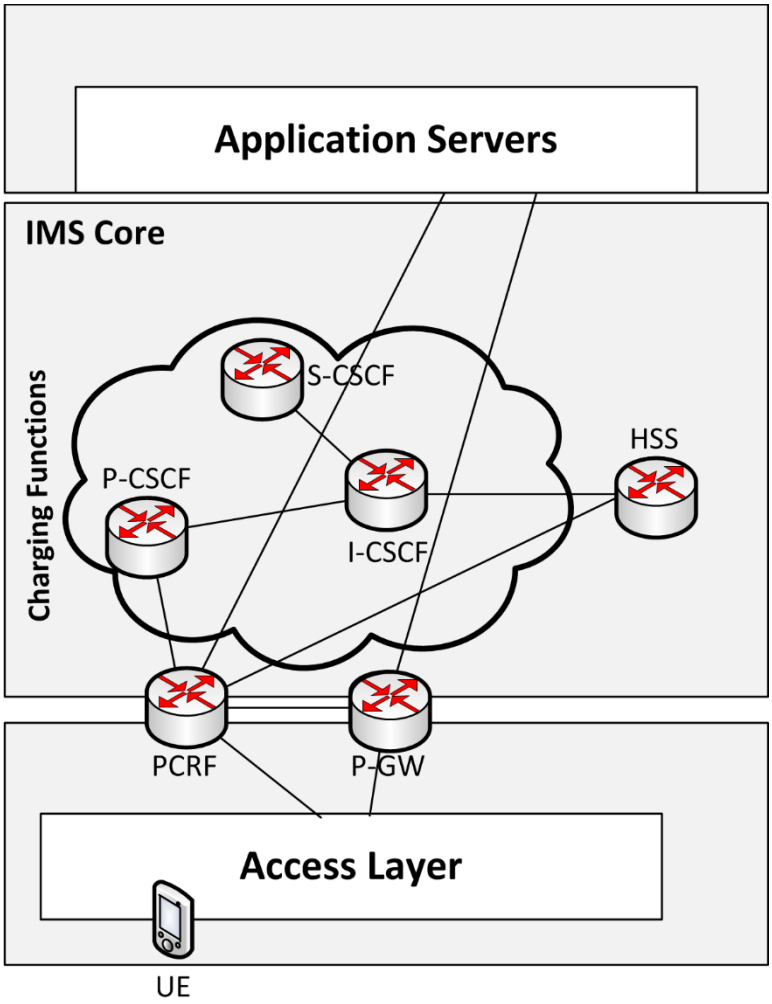


Figure 2.11 IMS Charging Functions

The performance of the IMS is affected by the individual performance of each entity inside it. The P-CSCF is the first entity that interacts with the User Element (UE) and forwards the SIP messages to other control functions in the IMS. It is also used to apply security or compression algorithms to the received traffic in addition to monitoring quality of service and bandwidth management. The I-CSCF is an SIP server that is assigned by the HSS to the user when it is required by the I-CSCF. The S-CSCF has interfaces with the HSS to perform session control, download the user profile information, assign the Application Server (AS) for further services, and to enforce the operator policy control. The SIP AS has an SIP interface with the S-CSCF, and is used to host specific IMS services. There are also gateways that function as interfaces with other domains.

2.5.1.1 IMS Registration Performance

In the IMS registration, the user requests authorization to gain access to the IMS services. Figure 2.12 shows the registration steps in IMS. First, the UE sends a SIP Register request to the P-CSCF, which forwards it to the I-CSCF. Then the I-CSCF sends a User Authentication Request (UAR) to the Home Subscriber Station (HSS), using the Diameter protocol, to check

the user profile for authorization process and to determine the S-CSCF address allocated to the user. Then, the HSS replies with the User Authentication Answer (UAA) to the I-CSCF and authorizes the user. After retrieving the S-CSCF details from the HSS, the I-CSCF forwards the SIP Registration request to the S-CSCF, which in turn sends a Multimedia Authentication Request (MAR) to download user authentication data to the HSS. The HSS then replies with the Multimedia Authentication Answer (MAA) to the S-CSCF, which in turn replies to the user using a SIP 401 Unauthorized response that embed an authentication challenge within it for the UE. The UE then generates another SIP Register request following the same steps as described earlier. This time the authentication process finishes and the S-CSCF sends a Server Assignment Request (SAR) to the HSS, using the Diameter protocol, which replies with a Server Assignment Answer (SAA) using the same protocol. Finally, the S-CSCF sends a SIP 200OK message to the UE so as to complete the registration process.

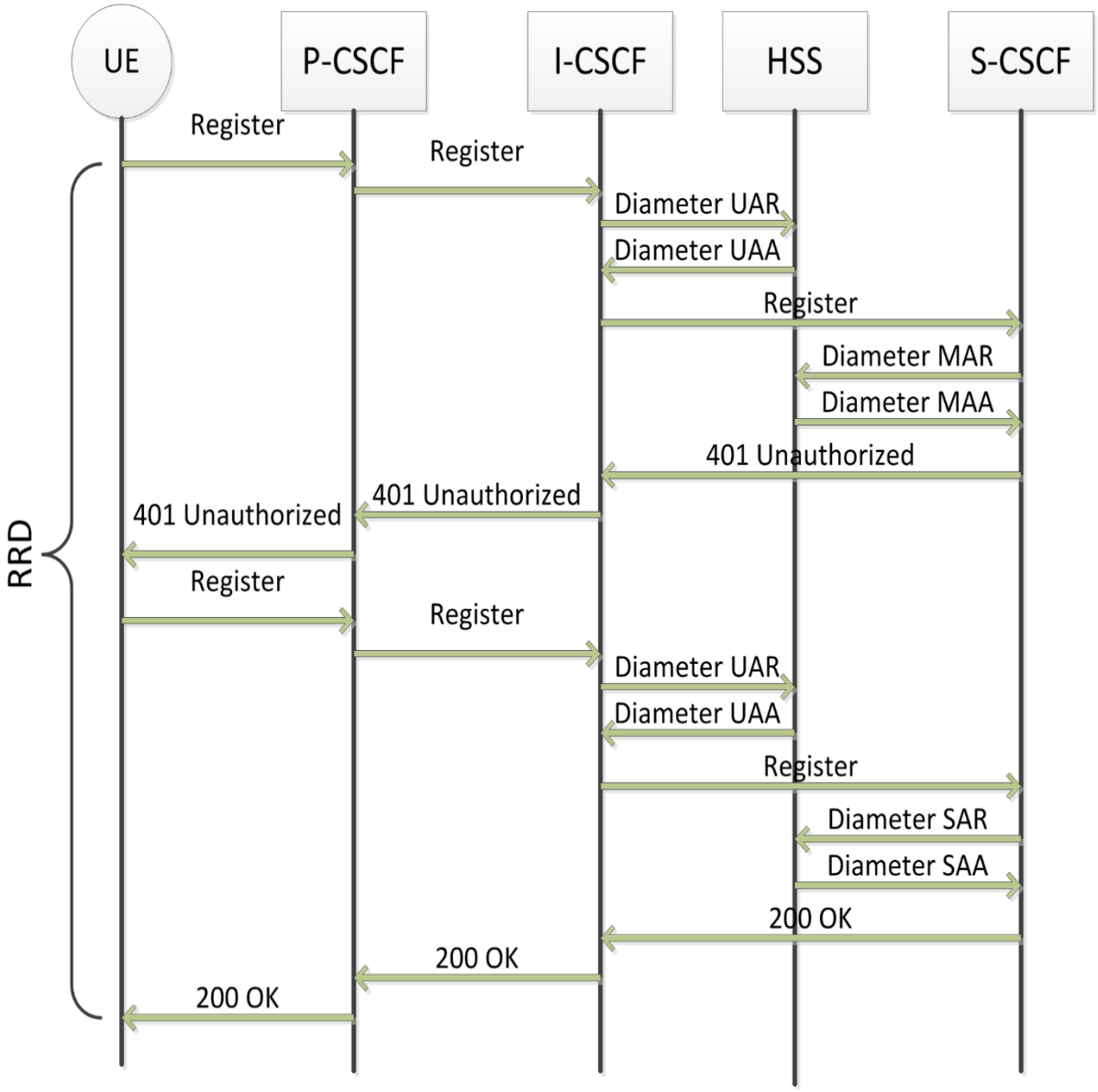


Figure 2.12 IMS Registration Process

As mentioned before, the Registration Request Delay (RRD) is the time needed for the registration process to complete. In this case, that is the time difference between the first SIP Register message sent by the UE and the time the SIP 200 OK message is received. Figure 2.13, shows the transition diagram of the IMS Registration process.

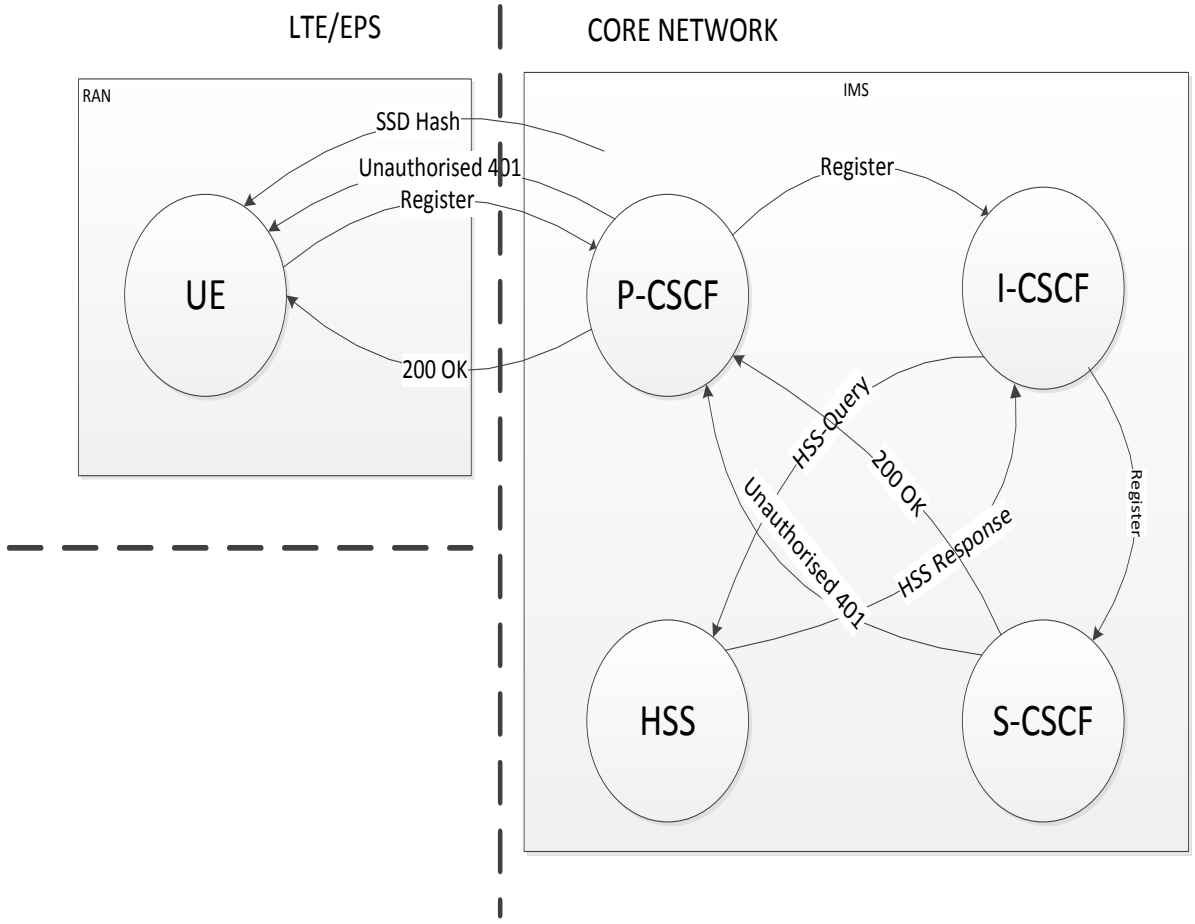


Figure 2.13 Transition Diagram of IMS Registration Process

During the registration process, the end user needs to be subscribed to the IMS in order to be authorized to make calls and initiate other services instantly during the call. The registration process is needed before initiating calls (see figure 2.14) and reregistration is needed for the SIP device to re-authenticate itself after receiving a notification from the SIP server. This is required in order to avoid compromising the SIP device.

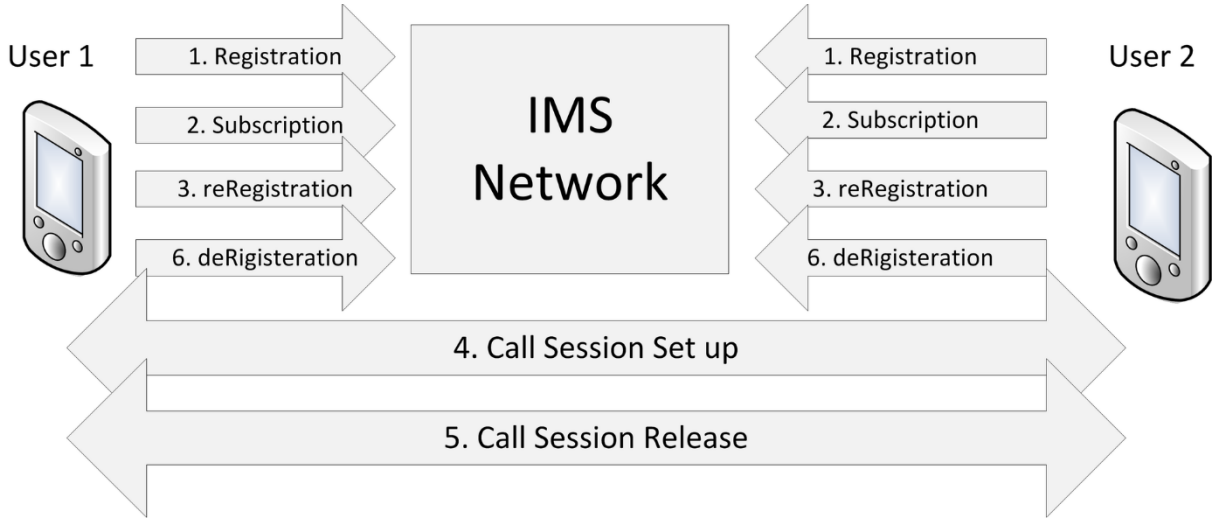


Figure 2.14 User Registration with IMS

The delay in the registration process is one of the SIP performance metrics and criteria that was defined by IETF in RFC 6076 (D. Malas 2011), due to the previous lack of a SIP benchmarking criteria to define the baseline performance of SIP signalling. RFC 6076 defines the performance metrics for the SIP in VoIP applications in order to provide key performance indicators and Service Level Agreement (SLA) indicators for best network resources utilization and best end user Quality of Experience (QoE). The main metrics defined in RFC 6076 are the Registration Request Delay (RRD), Ineffective Registration Attempts (IRA), Session Request Delay (SRD), and Session Disconnect Delay (SDD). RRD is the time needed for the user to finish the registration process successfully and plays a major role in the overall QoS.

In addition to the delay in the registration process, the rate at which registration messages are sent to the server is of great importance. This is especially true in mission critical operations where the number of users or devices that request access to the SIP/IMS server may be high enough to effect or even deteriorates the performance of the service. Different real-world scenarios for registering users/devices over LTE have been presented by other researchers. For example, in one scenario, an Airbus A380 carrying 1,000 passengers and landed at Heathrow airport was considered, with each passenger initiating IMS Registrations over the LTE at a rate of ten registrations per second. In the second scenario, a regional power outage is restored at a big city which causes a flurry of IMS registration requests to the network, 100 registrations per second is expected in this scenario (James Rankin 2014). In both scenarios, it is important to handle all the requests to successfully finish the registration process within acceptable time limits regardless of the received traffic especially over a mission critical system that is by nature should be highly available, resilient and reliable.

PoC is one of the mission critical services supported by IMS. It is designed to operate as a replica of Push To Talk (PTT) service in conventional dedicated mission critical communication systems such as TErristrial Trunked RAdio (TETRA) networks and P25. Therefore, the investigation of the IMS added delay in the registration process is of great importance for mission critical services which are less delay tolerant, especially if operating over a generic communication system that is not dedicated to mission critical communications.

Some studies that analysed IMS registration over WiMAX and 3G interworking architectures show that the access technology along with IMS design affect the registration signalling delay (Munir and Gordon-Ross 2010). In one study, a performance analysis of IMS signalling, including registration signalling, in multimedia networks was carried out. The study showed that the maximum signalling delay is affected by both the number of User Elements (UEs) and the available bandwidth available in the core network. However, the study was limited to a maximum of 400 users (Nader F. Mir 2012). Other researchers have proposed a general IMS registration protocol for wireless networks in order to reduce the registration delay by combining both the registration processes with the access technology and IMS network (Díaz-Sánchez, Proserpio et al. 2009). A reregistration procedure modification was proposed in order to reduce the transmission delay in the IMS (Farahbaksh R. 2007). Reregistration is needed either in order to refresh the registration state of the user or to update the user registration profile due to change in user capabilities. Moreover, when the impact of SIP signalling load over IMS was investigated, it was found that the SIP signalling load, including registration, has a significant impact on the Key Performance Indicators (KPI) for IMS (Jiri Hosek 2004).

This paper investigates the delay in the registration process introduced by the IMS for increasing numbers of UEs. Moreover, it also articulates the effect of increasing the load over the KPI and related QoE.

2.6 SIP KEY PERFORMANCE INDICATORS

A set of SIP KPIs is defined by 3GPP such as the Registration Request Delay (RRD) which is the difference between time of final successful registration arrival and the time of sending the registration request (3GPP 2006). Ineffective Registration Attempts (IRA) is a metric that indicates the ratio of unsuccessful registration attempts to the total number of registration requests sent. Session Request Delay (SRD) is the time difference between the reception of status indicative response and the time of sending the SIP invite message. Session Disconnect Delay (SDD) is the time difference between sending the SIP Bye message and the reception of confirmation or timeout message. The Session Duration Time (SDT) is the time interval between the receptions of the 200OK SIP message following the INVITE Request and the receipt of the BYE timeout message at the originating UE. The Session Establishment Ratio (SER) defines the ability of terminating the User Agent or Proxy Server during the session establishment. The Session Establishment Effectiveness Ratio (SEER) is the ratio of SIP Invite Requests that end with a 200OK response from the terminating side and the number of SIP INVITE Requests that end with a response type other than a 200OK. The Ineffective Session attempts (ISAs) is the ratio of failed session setup requests and the total number of session requests. Finally, the Session Completion Ratio (SCR) is the ratio of successful completed sessions and the total number of sessions. This paper, considers both the RRD and IRA as KPI metrics that reflect the performance of the SIP Registration process.

2.7 IMS KEY PERFORMANCE INDICATORS

The criteria for carrying out IMS performance evaluation using testbeds has been defined by the European Telecommunications Standards Institute (see their document (ETSI 2012)). Moreover, 3GPP in their technical standard (3GPP 2012) have defined a set of KPIs for IMS that can be classified in to three categories:

1. Accessibility KPIs: which are a set of metrics that reflect the accessibility of the IMS for the users. They are:

I. Initial Registration Success Rate (IRSR) of S-CSCF. This reflects the accessibility performance provided by the IMS and measures the percentage of attempted user initial registrations (UR.AttInitReg) that were successful. (UR.SuccInitReg). The percentage is calculated as shown in equation (2.1).

$$IRSR = \frac{\sum_{s-cscf} UR.SuccInitReg}{\sum_{s-cscf} UR.AttInitReg} * 100\% \qquad (2.1)$$

II. Mean Session Setup Time: measures the mean time for session setup starting from when the SIP Invite message is sent to the P-CSCF until the reception of the 200OK.

III. Session Establishment Success Rate: is calculated by the I-CSCF using two values: the ratio of successful session initiation requests to the total number of session initiation attempts; the ratio of successful session termination attempts compared to the total session termination attempts. Both shown in equations (2.2) and (2.3) respectively.

$$SESR\_Orig = \frac{\sum_{Type} SC.SuccSessionOrig.type}{SC.AttSessionOrig} \qquad (2.2)$$

$$SESR\_Term = \frac{\sum_{Type} SC.SuccSessionTerm.type}{SC.AttSessionTerm} \qquad (2.3)$$

Type is for SIP 180 and SIP 200OK message types

IV. Third Party Registration Success Rate: This KPI is obtained by dividing the number of successful third party registration procedures by the attempted third party registration procedures as shown in equation (2.4).

$$TPRSR = \frac{\sum_{s-cscf} UR.Succ3rdPartyReg}{\sum_{s-cscf} UR.Att3rdPartyReg} * 100\% \qquad (2.4)$$

V. Re-registration Success Rate of S-CSCF: the ratio of successful re-registrations of S-CSCF to the total number of attempted re-registrations as shown in equation (2.5).

$$RRSR = \frac{\sum_{s-cscf} UR.SuccReReg}{\sum_{s-cscf} UR.AttReReg} * 100\% \qquad (2.5)$$

VI. Mean Session Setup Time Originated from IMS (MSSTOI): the mean setup time of successful originated calls from the IMS.

2. Retain-ability KPI: it has one value, the Call Drop Rate of IMS Sessions, which is calculated as the ratio of dropped sessions to the number of successful session established as shown in equation (2.6).

$$\text{SEDR} = \frac{\text{SC.DroppedSession}}{\sum_{type} \text{SC.SuccSession.}type} \qquad (2.6)$$

3. Utilization KPI: it has only one value, the Mean Session Utilization (MSU), which is calculated as the percentage of the mean number of simultaneous online answered sessions compared to the maximum number of sessions provided by the IMS network as shown in equation (2.7).

$$\text{MSU} = \frac{\text{SC.NbrSimulAnsSessionMean}}{\text{Capacity}} \times 100\% \qquad (2.7)$$

The QoE, which measures the quality of the service as perceived by end users, is affected by the KPI values of both the SIP and IMS. In order to improve the QoE, the related KPI values need to be continuously monitored and improved. The KPI measure, if combined together, will help in determining the failure points in the IMS network and be indicative of how to improve things so as to avoid the bottleneck and single point failure in case of traffic overload scenarios.

The KPI may also be used as a Load Detection Function (LDF) (3GPP 2013). The LDF may be used to support load balancing for the S-CSCF and, therefore, reduce the service request latency.

2.8 VOICE APPLICATIONS

There are different types of applications and data formats used in mission critical communication system. These include voice, video, and push messages. Some of these applications are termed “killer applications” due to their bandwidth requirements especially at the core and access point side, which are considered bottlenecks in the overall system. Another application, Push To Talk (PTT) enables the caller to push a button to initiate the sending of a voice burst to the callee side. This particular application uses half-duplex channels to save bandwidth and minimize the need for extra bandwidth .There is a need for addressing users scalability concerns that may emerge due to two-way or full-duplex communication methods which will be a challenge for future broadband mobile communication systems if PTT over Cellular (PoC) are adopted in the future. However the adoption of PoC introduced additional delays especially if it was applied over LTE which experience more delays that need to be analysed (Lu 2012)

Voice over IP (VoIP) is used for the delivery of voice bursts over IP networks. Due to the emergence of All-IP-Networks, VoIP has gained significant importance as the best voice streaming solution for voice calls. Voice Codec is used at the subscriber side to convert the

analogue voice waves to digital pulses and vice versa. There are different codec types based on the selected sampling rate, data rate, and implemented compression algorithm. In order to determine the bandwidth requirements of VoIP connections, there is a need to identify common VoIP Codecs and their associated characteristics (table 2.4 shows the Bandwidth Requirements of different Codecs (Ali, Vassilaras et al. 2009). VoIP services are considered bandwidth killing applications for mobile broadband Access Technologies that provide end users with services similar to those of DSL providers in the traditional wired networks. However, the wireless access mechanism introduces additional challenges and constraints on the data rate as well as delay requirements. Therefore, the need for a bandwidth requirements study for VoIP was needed to ensure that the mission critical service that use VoIP operate as expected and to determine how best to increase the scalability of the networks within acceptable QoS levels.

Table 2.4 VoIP total Data Rate for Different Codecs (Ali et al., 2009)

Codec	G.711	G.729	G.723.1[A]	G.722.1[B]	G.722.1[C]	G.722.1C[D]	EVRC[E]	EVRC[F]	EVRC[G]	AMR[H]	iLBC[I]
Sample Time (ms)	10	10	30	20	20	20	20	20	20	20	20
Frame Size[J]	640	80	158	480	640	960	172	80	16	95	303
Packets Per Second	100	100	33	50	50	50	50	50	50	50	50
IP,RTP,UDP headers	320	320	320	320	320	320	320	320	320	320	320
Ethernet Header[K]	176	176	176	176	176	176	176	176	176	176	176
Total Size	1136	576	654	976	1136	1456	668	576	512	591	799
Ethernet CS data rate[L]	113.6	57.6	21.80	48.8	56.80	72.8	33.4	28.8	25.6	29.5	39.9
IP CS data rate[M]	96	40	15.9	40	48	64	24.6	20	16.8	20.7	31.1

A. 5.3 Kbps B. 24 Kbps C. 32 Kbps D. 48 Kbps E. Rate 1 F. Rate 1/2 G. Rate 1/8 H. 4.75 Kbps I. 15.2 Kbps
J. All sizes measured in bits K. Including the CRC header and 802.1Q header L.Calculated in Kbps M. calculated in Kbps

2.9 MULTIMEDIA SERVICES IN MISSION CRITICAL SYSTEMS CHALLENGES

Based on the earlier discussion, there are two different deployment options for the mission critical systems, either to have a dedicated system or a commercial system. Unfortunately, there are drawbacks for both options and many trade-offs that control the design and deployment process. What is important to note is that at the nascencey of dedicated mission critical systems, voice and simple text messages were the only services used by PPDR for their critical missions. Simply during the early stages of the evolution of mobile communication networks operators only thought of providing voice communication between users. Hence, when dedicated mission critical communications were deployed, the standardization bodies designed the whole system only for voice and low data rate services. However, due to the exponential increase of mission critical data used to support different services and tasks, the need for broadband data along with the traditional basic voice service has become vital for the mission critical communication systems.

Due to the increased demands for more bandwidth, the need for an access technology that is capable of providing broadband connectivity and scalability, along with broadband network

support in the core network, is of crucial importance for any mission critical communication system.

2.9.1 Access Technology Challenges

As discussed earlier, there is an increasing demand for more bandwidth for mission critical services. This implies that systems not only need to satisfy the general requirements of mission critical services, but also a more scalable system to serve more users. Moreover, the users' broadband services should be running as it should be during the crises and worst case scenarios. It is important to note that the mission critical communication system is an all-time-available system that should be design for the worst case ever scenarios not for the normal daily routine PPDR tasks.

Increasing the bandwidth of the system is not a simple task in the physical layer of any communication system especially that there are multiple factors and trade-offs that complicates the evolution process of any communications system. From a business perspective, mobile network operators need to utilize every Hertz of the invested licensed frequency band to ensure the maximum profit. Technically speaking, the frequency spectrum should be utilized so to achieve the highest possible bandwidth to support the broadband services. Hence, the need to deploy methods for the best frequency exploitations such as: adaptive modulations, Multiple Input Multiple Output (MIMO) Antennas, Orthogonal Frequency Division Multiple Access (OFDMA), Code Division Multiple Access (CDMA), Frequency Division Duplex (FDD), Time Division Duplex (TDD) and many other technologies in both physical and MAC layers.

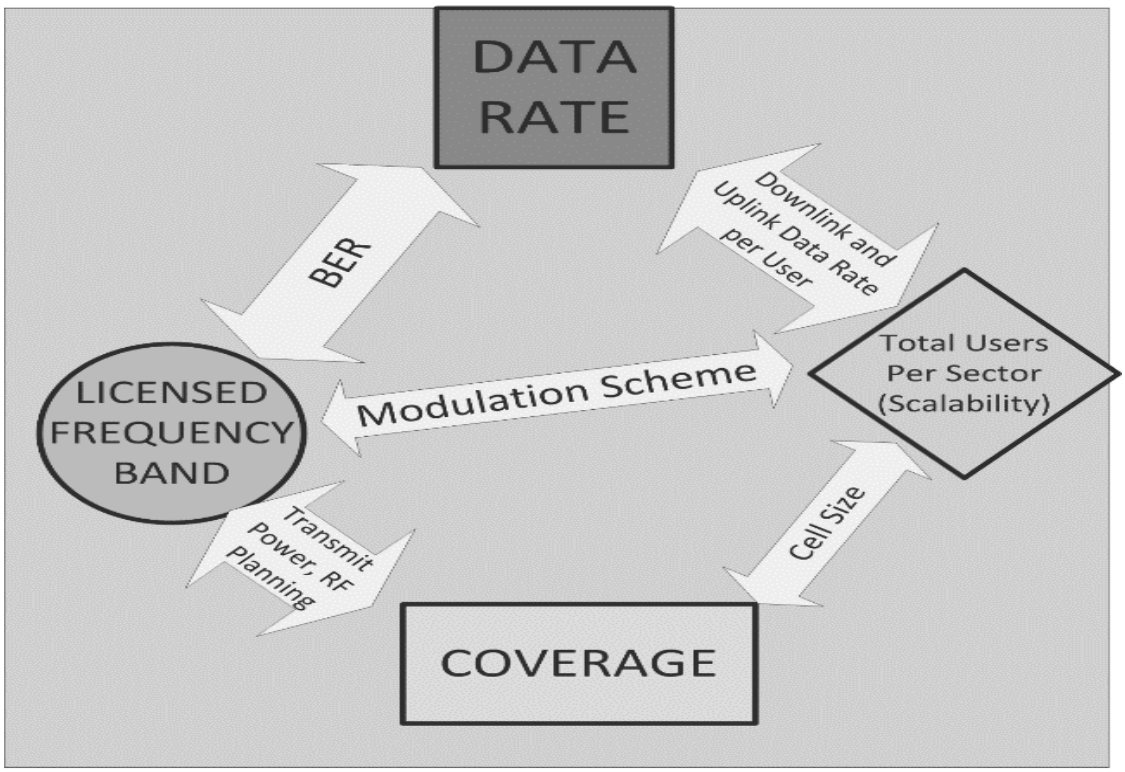


Figure 2.15 physical and MAC layer challenges

As shown in figure 15, There are trade-offs between the frequency, transmission power, coverage area, and cell size. In addition, there is trade-off between the frequency and maximum

system throughput based on the Bit Error Rate value. Finally, the system throughput controls the downlink and uplink capacity of the resource blocks which limits the maximum number of simultaneous users and the scalability of the whole system.

Current operational 4G mobile communications networks are widely deployed using the Long Term Evolution (LTE) standard articulated by 3GPP. While the LTE standard can be part of a public safety system, it is important to note that previously existing trade-offs still cause issues within the access technology abstract. However, the LTE or other technologies dedicated technologies can provides reasonable solutions in the physical and MAC layers to maximize the capacity of the system.

2.9.2 System Resources Sharing Challenge

Another challenge is using commercial mobile communication networks as a public safety communication system. As mentioned before, there are two options - either a dedicated system or commercial system for the mission critical services. Clearly, it is a significant challenge to enable a general-purpose public commercial communication system to work as a replica of a dedicated public safety communication system with all the mission critical services and strict requirements that have to be satisfied. On the one hand, using the dedicated system will enable all the communication system resources to be exclusively available to the PPDR users. On the other hand, for the commercial public safety communication system, the resources will be shared between the PPDR and public commercial mobile users. Therefore, for the public communication system the need to have a distinction between the mission critical services and other non-mission critical services in terms of service priority handling and resource allocation. This ensures that the system will have enough resources and thus be available for the PPDR users whenever and wherever needed.

Part of this study will focus on the cross layer communications between different system abstracts to ensure that calls, such as mission critical emergency call signalling, with higher priority class, will pass through the different entities of the system toward the end user and at the same time satisfying all the mission critical service requirements. In addition to that, the need to have a pre-emptive scheduling algorithm to decide which type of calls need to be dropped to flush out the allocated resources in cases where system capacity may reach to its maximum tolerable limits. Although the deployment of pre-emptive scheduling algorithms is out of the scope of this study, however a thorough evaluation of overall system capacity with different scenarios that combine the number of users along with the priority class of the service will be analysed and investigated.

2.9.3 System Architecture Design Challenge

Mission critical communication systems are composed of different entities and interfaces to ensure end-to-end connectivity between end users and overall system integrity. End users may be connected to different network domains and access technologies, hence the need for interfaces and gateways that provide transparent connectivity between the different domains. In addition, there is a need for standardized interfaces that connect devices regardless of the

manufacturing vendors. Finally, there is a need for a data plane and control plane protocols that ensure the connectivity and data delivery between end users.

Figure 2.16 shows the general layered architecture model. Regardless of the details of any mission critical communication system, all systems share the same hierarchal architectural structure. End users are the ones holding the mobile device that is running the applications that interact with the required services through the available access technology. The access technology that is being used to reach the core network can either be a single predetermined access technology or multiple access technologies. Clearly, having more than one supported access technology to access the core network will support the reliability requirement of requested mission critical service.

Regardless of the access technology that is being used, the need to have a common platform working between the access technology and the mission critical application servers is a necessity for two reasons. Firstly, the platform needs to work as intermediary between the control signalling and data streaming signalling domains. Secondly, the platform is needed to provide seamless connectivity for the end-users along with the application servers regardless of the access technology being used. Moreover, it should be noted that there is no clear distinction between the data domain and control domain signalling protocols.

In this study, a framework for the mission critical communication system will be proposed and explained. The different layers and interfaces for both data plane and control plane will be demonstrated. The layered architecture model of the system and the details for each layer will be investigated and demonstrated.

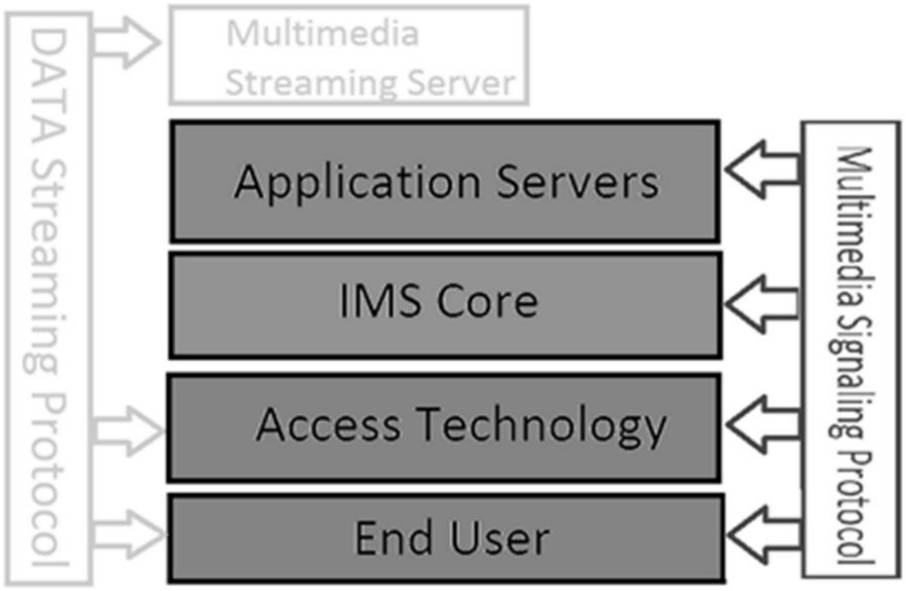


Figure 2.16 Layered Abstract Model

2.9.4 End-to-End QoS Challenge

Having different abstracts, in hierarchal manner for the architecture of the system as described previously, will introduce not only integration challenges but also a cross layer service performance optimization challenges. This is crucial in order to improve the quality of service

for the whole end-to-end mission critical communication system and to satisfy the mission critical requirements mentioned previously.

2.9.5 System performance modelling and validation challenge

Modelling a system and simulating its behaviour, while changing a set of dynamic parameters, is useful in providing an understanding of its overall functionality and responses. Verification and validation are essential parts of the readability and availability testing process.

The verification process of a mission critical system starts with a model design (or framework) analysis. The system design needs to be checked against all specifications before the system is implemented and built. For a newly proposed system, the technical reports and the standardization bodies in the field are responsible for setting the guidelines and set of specifications need to be followed by system engineers. This helps ensure a more realistic and accurate model. The lack of finalised standards imposes many challenges in the verification and validation steps.

2.9.5 Proposed Solutions

The challenges outlined in the previous section highlight the need for a further study to focus on the cross layer communications between different system abstracts. The result of this study should help ensure that calls, such as mission critical emergency calls, with higher priority class, pass through the different entities of the system toward the end user while at the same time satisfying all mission critical service requirements. In addition, there is a need to develop a pre-emptive scheduling algorithm capable of deciding which type of calls need to be dropped to free up allocated resources at times when system capacity is being exceeded. To facilitate the development of such a pre-emptive scheduling algorithm, a thorough evaluation of overall system capacity with different scenarios that combine a number of users along with the priority class of the service will need to be undertaken.

The proposed study will need to develop a framework for the mission critical communication system. Moreover, the different layers and interfaces for both data plane and control plane need to be demonstrated. Furthermore, the layered architecture model of the system and the details for each layer need to be investigated and clarified.

A cross layer optimization and integration of the signalling used in both control and data domains need to be proposed as part of a general framework. Finally, in order to test the proposed framework and the overall proposed system architecture in addition to validating the proposed cross layer optimization, the whole system need to be modelled using simulation tools. The Quality of Service performance measures need to be compared with the performance metrics of a testbed for the mission critical system that will be used as a benchmark.

One of the main objectives for this study is to propose a cross layer optimization solutions between both the interfaces and protocols running in the mission critical system to help facilitate the enhancement of the overall performance and its associated metrics.

2.10 SUMMARY

This chapter articulates some of the many challenges that need to be addressed with respect to current commercial mobile communication systems. The main goal is migrating to a generic system that is better in terms of performance than current dedicated mission critical communication systems and at the same time more scalable to serve increasing number of commercial mobile users. The performance of SIP and IMS, as a core part of the generic proposed system, has significant implication on the system's overall performance. The registration process, as an example, is considered a good way to measure the system response time as most of the system entities are involved during the registration process. Similar to the registration, there are a set of KPIs for both the IMS and SIP that can be considered as a reference points to measure the system performance at different layers and abstracts of the end-to-end system. The challenges and requirements that need to be addressed satisfied in any proposed alternative have also been articulated.

Chapter 3: LITERATURE STUDY AND RELATED WORK

3.1 INTRODUCTION

This chapter will present the recent and related enhancement in literature that are related to the topics covered partially or in depth in this thesis. The chapter is split into four parts; SIP performance, IMS, Mission Critical Systems performance, Literature and related studies. Then it will introduce to the reader the current challenges that has been extracted from the literature study and others related work.

3.2 SIP PERFORMANCE

The Session Initiation Protocol (SIP) is the considered the core entity for multimedia communication system. It is designed to handle the SIP generated messages by both the client and server sides, User Agent Client (UAC) and User Agent Server (UAS) exchange different types of SIP messages during the call setup, modification and termination. To avoid performance bottleneck, the SIP server performance need to be investigated and enhanced to ensure that it will not interrupt the already established and running SIP sessions or block new session establishment and termination whenever needed.

In this section, the enhancements in SIP standard performance and SIP performance related studies will be presented. Understanding SIP performance is considered essential for understanding the overall multimedia system communication performance. IMS for example, is composed of many entities that communicate with each other using SIP signalling.

In (Krishnamurthy and Rouskas 2015) the authors presented the impact of the scheduler settings over the overall performance of multithreaded servers. The operating system and the hardware architecture of the machines that host the SIP server has a great impact over the SIP performance. Service scheduling for multithreaded systems for example is controlled via process scheduler which ensures that the multi process and threads within the SIP server are served fairly and all the available resources are utilised appropriately. The Complete Fair Scheduler (CFS), for example, is used in Linux based systems as a scheduling policy that schedules the threads need to be served by the CPU based on certain fairness criteria. The study relied on using certain parameters to alter the default operation of the CPS for better scheduling performance and therefore better SIP server performance.

In (Krishnamurthy and Rouskas 2013) the authors evaluated the SIP proxy server performance via packet level measurements. A testbed that emulate the SIP messages generation and SIP server was introduced. Both methodological mass data was collected and compared against a queuing model for SIP Proxy server, the main objective of the experiment is to get a precise measurement of SIP messages processing and serving time spent in the kernel and SIP layer.

Authors of (Krishnamurthy and Rouskas 2013) used an open source SIP server (OpenSIP) over Linux operating system along with open source SIP packet generator (SIPp) to emulate the SIP user-server communication signalling. The source code for both the generator and the OpenSIP along with the Operating System Kernel source code were modified to record the SIP message type and timestamp of the exact time it was received at different layers. This will give more details of the exact time certain SIP packet has taken during each layer abstract including the SIP layer processing time.

In (Krishnamurthy and Rouskas 2016) the authors evaluated and investigated the impact of using load-balance scheduling algorithms over the overall performance of OpenSIPS (Open Source SIP Proxy Server) which is an open source SIP server that operates over multi-processor Linux based Operating Systems (OS).

3.3 SIP PERFORMANCE METRICS

The performance characteristics of SIP elements need to be studied and identified to decide the feasibility and service potential of certain setup and the cost of proposing alternative solutions. For certain services in mission critical systems, such as Push-to-talk Service, the time constraints is so strict that the end-to-end communication need to setup and ready for both call parties in order of 10's or few hundreds of mille seconds. This enforces the intermediate nodes to not only having the ability of processing the SIP messages in a timely manner, but also to be designed to comply with the all SIP end-to-end performance metrics and measures to fully support different types of multimedia services.

In this section, different performance metrics found in the literature for SIP will be introduced and briefly discussed. In (Cortes, Ensor et al. 2004) the authors have subdivided the SIP performance metrics into two main categories (Protocol Related Metrics and Implementation Related Metrics). The SIP message structure impose a significant load on the message handler routine. The message syntax complexities are listed below and explained briefly afterwards:

- Text-Based Message
- Unordered Headers
- Verbose Headers
- Variable Size Messages
- Case-insensitive Keywords
- Long-lived messages
- Message extension

SIP Messages are text based and its structure has a command line and three or more headers, all SIP nodes need to perform string operations over the header values and command text. The headers can be out of order and in multiple times and separated away from each other in the message structure, this introduces complexity in the parser operation which in certain occasions need to read the entire message before deciding the location and the interpretation of multi occurrence headers. Each header and parameter in the SIP message has a header name and parameter name that work as a tag for the parser to easily find the value and parse it. This adds

more load on the parser and require an efficient parser function. The variable message size, which is due to variable header value sizes, which impacts the CPU and memory performance in each SIP node. Except for the command line, all the headers in the SIP message are case insensitive, the parser need to recognise all different forms of the keywords to be able to correctly parse it. The long-lived messages is needed for certain SIP nodes, this implies that for state-full proxies may need to hold the SIP message for longer period to keep track of the call state. The message extension gives the SIP message a dynamic structure by adding up to 33 new header extensions, but this adds additional complexity over the parser performance as it need to expand the keyword table to include the additional extensions.

The other category mentioned in (Cortes, Ensor et al. 2004) is the Implementation-related Metrics. In the SIP node, there are too many processes occur as a reaction of the reception of one single SIP message, such parallel process are executed using multiple threads using multi-threaded multi-tasking processors. The tasks need to be accomplished for one SIP message need to be executed in parallel in order to utilise the multithreaded parallel execution power of the CPU. Reading a message header, parsing a specific field value, replying or forwarding a message are all examples of different tasks that can be grouped into single or multiple stages based on the agreed implementation architecture.

Each stage can be executed via certain pre-specified number of threads, once the stage execution finishes, the next stage starts with its own set of processes and threads. The parallel execution of threads implies that there will be a need to access shared resources at the same time, this rises the need to enforce a data consistency and coherency policy to ensure correct execution of the entire stage. Therefore, concurrency control primitives are needed to protect the shared information. Mutex Locks are used to protect the shared data by enabling one thread only to control the Mutex Lock at a time, once another thread try to use an already booked Mutex Lock, it need to wait until the lock is released by the running thread before passing it to the waiting thread. The waiting time can be utilised by executing other free threads via the multithreaded processor. The waiting thread will spin for a number of CPU execution cycles in hope that the running thread will finish it execution and release the lock, if not, then the CPU will schedule the waiting thread in a waiting list to be executed later. Clearly, the cost of switch execution threads is much higher than the spinning time. According to the previous discussion there are two main Implementation Related Metrics for SIP collected by the OS:

- Number of Spinning Locks per second
- Number of Context Switching per second

The first metric reflect the number of collisions between threads while waiting to get a free lock. Whereas the second metric counts the number of times the OS reschedules the thread or process for another cycle of execution after being held in a waiting queue.

3.4 END-TO-END VS SINGLE HOP SIP PERFORMANCE

QoE and QoS is measured by either subjective or objective methods. It is important to understand that the user perception of a service is not only affected by the quality of the multimedia data stream, such as audio and video, it gets from the other side of the network, but

also depends on the Signalling quality of the control message being sent to establish, configure, or tear down the connection between end users. SIP Signalling performance metrics have been well defined in literature and even standardised by the 3GPP community. In literature, the SIP performance metrics can be categorised based on the scope of transmission and performance measurement, SIP performance can be either measured from the end-to-end based on end users perspective or as a single transaction between two nodes. According to this, we get two SIP performance metric categories; an End-to-end SIP Performance Metrics (Husić, Bajrić et al. 2012), and a Single Hop Performance Metrics (Happenhofer, Egger et al. 2010, Happenhofer and Reichl 2010).

In (Happenhofer, Egger et al. 2010, Happenhofer and Reichl 2010) the authors define the concept of Quality of SIP Signalling (QoSg) in which they used the IP performance Metrics (IPPM) to study the smallest building block in the SIP Signalling (Single SIP transaction). In (Happenhofer, Egger et al. 2010) the quality of signalling for non-INVITE messages was studied, the non-INVITE signalling are used for the Registration, SIP user agents capabilities exchange, Presence information exchange, and instant message delivery, update the session parameters using UPDATE method, or terminating a call using BYE method.

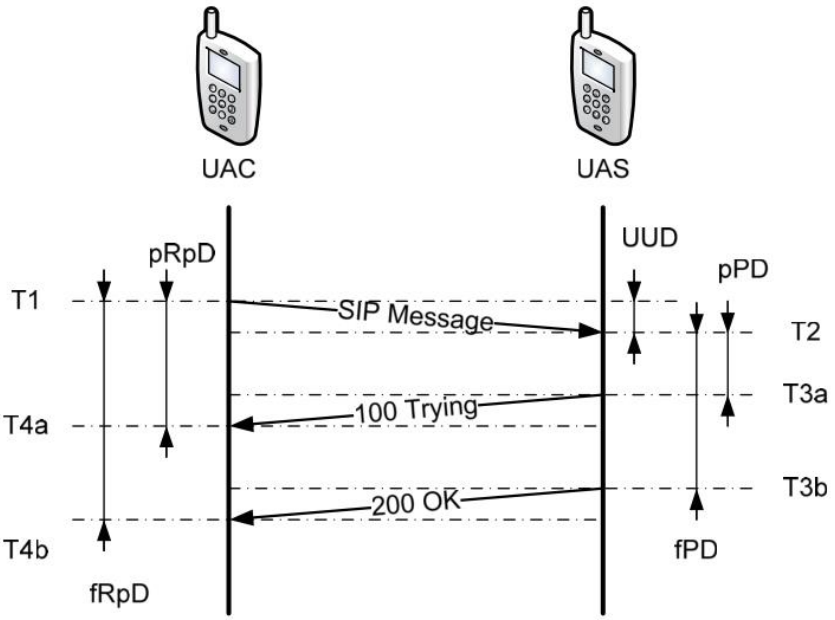


Figure 3.1 signalling timestamps

As shown in figure 3.1 there are different timestamps set at the both the client and server sides. The timestamps are used to calculate three main metrics that defined the performance of single SIP transaction:

- User to User Delay (UUD): which is simply the time interval from sending a request until it arrives at destination.
- Processing Delay (PD): represent the time the UAS needs to process the request and send the response message.
- Response Delay (RpD): is the time interval the UAC waits starting from sending the request till it gets a response back.

The success rate and the traffic of a transaction can be calculated as well, the request Transmits (RT) and responses transmits (RpT) reflect the rate at which either requests or responses are sent.

3.5 QUEUING MODELS FOR SIP PERFORMANCE

The session initiation protocol (SIP) is now the preferred signalling protocol for internet telephony due to its simplicity in both signalling and syntax along with debugging and compiling features. The need for characterizing the SIP signalling and the performance of SIP network is crucial to be able to expect it performance and reliability in real world scenarios. The performance study via modelling has been carried out via many researchers using stochastic processes, queuing and performance analysis that try to replicate the operation and performance of real physical systems in real time.

Analytical models help in evaluating the performance of a physical system, such as communication networks, to better understand the system characteristics that may affect the performance aspects. Performance metrics and parameters need to be defined to be able to benchmark the system performance and decide the needed enhancements. In this section, different performance models found in the literature will be presented.

In (Gurbani, Jagadeesan et al. 2005) the authors started with the first reliability and performance study of the SIP systems over current networking systems, the objective of the study is to come up with new model to characterise the SIP performance and determine new SIP performance metrics suitable for the internet domain as the old models only studied the circuit switched domain. The authors provided analytical models for the performance analysis of for SIP network, then they used the proposed model to analyse the performance the SIP performance with varying parameters such as the network delays, service rate and arrival rate. Then a reliability evaluation of SIP Network along with evaluating the lost calls in SIP networks, then they compared the results with the more reliable Public Switched Telephony Networks (PSTN) reliability and performance measures. The model assumptions are summarised in Table 3.1.

Table 3.1 Model Assumptions

INVITE message service time	1/ μ	Network type	Lossless
180 message service time	0.3/ μ	Call Arrival Rate	λ
2XX message service time	0.3/ μ	Station Queue Model	M/M/1
Non-2XX message service time	0.3/ μ	UAS waiting time between sending 180 Ringing and 200 OK messages	Zero
UAS service time for 180 Ringing Response message	0.7/ μ	Proxy Service time of 180 Ringing message	0.3/ μ
UAS service time for non-2XX Response message	0.5/ μ	Proxy Service time of 2xx message	0.3/ μ

Based on the assumptions mentioned, the Feed-forward model for the SIP Network is shown in figure 3.2.

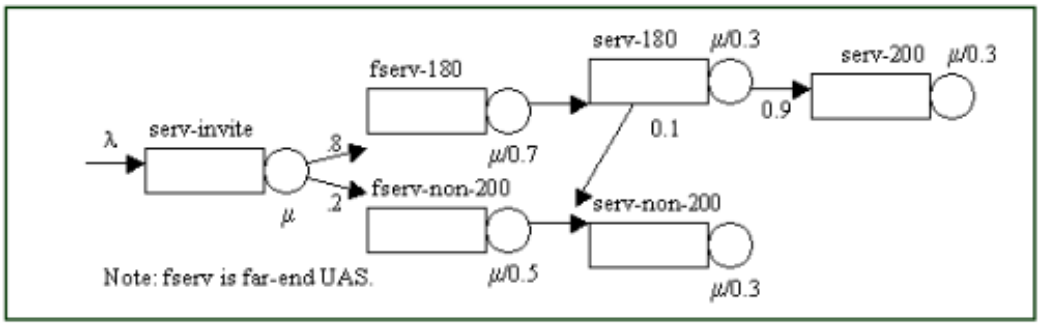


Figure 3.2 Queuing Model

The models shows that when the intermediate proxy forward the Invite message to the UAS, there will be an 80% probability that it will get a 180 Ringing message from UAS (Far end Server), otherwise it will get a non 2xx message from UAS with a probability of 20 %. The UAS meanwhile will serve the INVITE messages it got from the Proxy server with a service time of 0.7/ μ or 0.5/ μ whether it is replying with 180 Ringing or non-2xx response message respectively. Then, the proxy server will process the received 180 Ringing message with a service time of 0.3/ μ before either getting 200 OK message from UAS or a non 2xx message which indicate a failure in establishing a connection. Both (2xx and non 2xx) responses will be received by the proxy server with a probability of 90% and 10% respectively, implying that the end user will most likely answer the call with 90% success chance. If the call is answered, the proxy server will process the received 2xx message with a service time of 0.3/ μ and the call will be established. Otherwise, if the 2xx is not received or the session times out without getting any reply from UAS, then a non-2xx message will be sent back to the UAC with a service time of 0.3/ μ.

In (Subramanian and Dutta 2008), the SIP performance model proposed in (Gurbani, Jagadeesan et al. 2005) was emulated and the analytical results obtained from the model was compared against a real SIP proxy server experimental results. Finally, the authors proposed a less complex and more predictable M/D/1 queuing model instead of the M/M/1 model proposed in the earlier study. The authors validated the model with several set of experiments using real SIP proxy server and compared the collected data with the analytical results. The experiment used a local Linux based DNS server to avoid interference and unnecessary delays that may affect the SIP performance. The setup implemented with one UAC, one UAS, a Proxy server, an Ethernet switch and DNS server. SIP Express Router (SER) is installed as a proxy server, SIPp tool is installed on the UAS side and another instant of SIPp tool is installed at UAC side to generate calls at different rates. The authors compared between analytical and experimental collected data and it was noticed that there is a slight difference that increase to be significant with the increase of the call rate. Hence, they proposed a new M/D/1 model with fixed serving time as it was noticed during the experiment that the SIP server will need the same time to process all SIP messages types. The researchers found that the new model has less mean

number of jobs and mean serving time values of the system that are closer to the experimental results they got. Which makes the new proposed model more accurate.

In (Subramanian and Dutta 2009), the paper presented a more realistic and accurate model based on M/M/c queuing model compared with the earlier M/M/1 and M/D/1 models. The authors redesigned the SIP Proxy software using multithreading technique that is more advanced than the simple feed forward queuing models proposed before. The experiment was set using CISCO proxy server that process the INVITE requests and other SIP messages received from the transport layer in one queue differently based on the message type and in concurrent threads. This design reduces the SIP messages processing time and increase the arrival rate to the system.

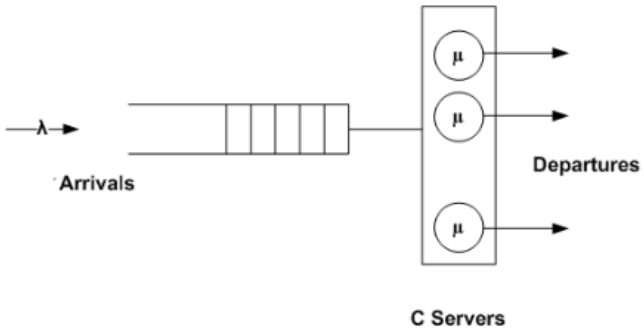


Figure 3.3 M/M/c Queuing Model

Based on the experimental and model analytical results, it was found that the M/M/c queuing model with one queue and multiple threads has a better prediction of server performance compared with the original M/M/1 model with multiple queues each with a single thread. Figure 3.3 shows the M/M/c model layout.

The results of (Subramanian and Dutta 2009) was further extended and validated in (Subramanian and Dutta 2010). The authors has studied additional performance characteristics for the SIP server and introduced additional performance measures such as queue size, CPU utilization, memory utilization, and generally accepted Call Hold Time (CHT). The results shows that the M/M/c model better scaled in term of adapting the increasing number of incoming calls. The server utilization and buffer occupancy values were close to the real experimental results using the proposed model.

3.6 IP MULTIMEDIA SUBSYSTEM

IP Multimedia Subsystems (IMS) supports the Quality of Service (QoS) signalling and negotiation of parameters. IMS depends heavily on SIP signalling between its entities. It is important to know that the SIP performance discussed in the previous section plays major role in IMS performance

Part of IMS overhead signalling was discussed in (Cortes, Ensor et al. 2004) in which the authors has studied the added SIP signalling complexity overhead introduced by IMS compared to a simple user agent back to back SIP server. The research focused it analysis on one IMS entity only, the Serving Call Session Control Function (S-CSCF), S-CSCF is considered the

core entity in IMS as it has interfaces with all other entities and connected directly to the Application Servers (ASs). The S-CSCF is involved in both user registration and call establishment counterparts. Once the SIP REGISTER message gets into the appropriate S-CSCF which in turns will process the request further based on the filter criteria in the user's subscriber data. Similarly, when the S-CSCF gets INVITE message and other call control messages, it will process it based on the user subscription data via checking the user registration status and enforcing basic service constraints. The S-CSCF will also check the user filter criteria and forward the message to the targeted AS, once the final reply is received from the AS, the S-CSCF will finally forward the original user SIP message to the destination user.

It is important to note that there are many SIP interactions that takes place before a user gets into its S-CSCF. Once a user is linked to S-CSCF, other interactions will take place as described before and shown in figure 3.4.

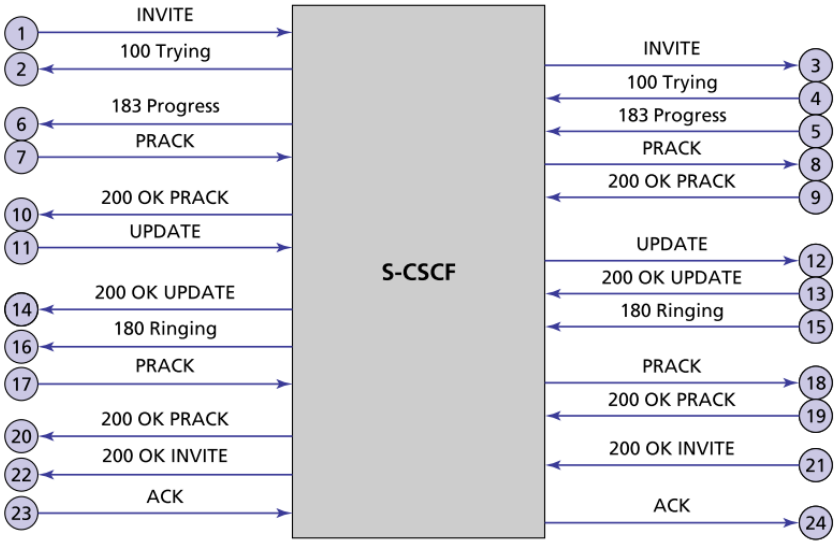


Figure 3.4 IMS interactions

The IMS baseline calls, without AS involvement as shown in figure 3.4 , implies that the IMS call setup SIP signalling is more complex, almost twice, than simple SIP server signalling, this is due to the additional media negotiation and additional reliability signalling. Adding one or more AS to the signalling overhead will even increase the complexity more. Based on (Cortes, Ensor et al. 2004), the number of SIP messages processing load will increase linearly with the number of Application Server, typical simple SIP server INVITE message without AS needs 6 transactions, using IMS without AS will need 14 incoming and outgoing SIP transactions. With added AS, the total number of messages will be 14+14n for incoming calls and another 14+14n for outgoing calls making the total 28+28n messages; where n is the number of Application Servers. In summary, the number of message load grows linearly with the number of application servers participating in the session.

Another comparison between the SIP and IMS performance was carried out by authors of (Kellokoski et. al. 2010) in which the call and messaging performance was compared and evaluated for both systems. The performance for both system was performed using the same configurations, hardware device specifications, and the same performance metrics. The results

shows that the IMS call initiation time takes 2 to 2.5 times the simple SIP call initiation period. Similarly, the successful message delivery time in IMS was slower than that in SIP but 99% of messages were delivered successfully within 100 ms.

Due to the overload introduced by IMS as explained before, many research was carried out to overcome the added overhead without compromising the basic functionality IMS was proposed for. In (Mishra et. al. 2014) the authors have proposed a model to reduce the overload and retransmissions in IMS. SIP retransmissions in IMS occurs due to the fact that SIP messages are being sent over User Datagram Protocol (UDP), which is an unreliable transport protocol. The retransmissions along with the already existing SIP overhead in IMS networks introduces an additional overhead in the system. The authors in (Mishra, Dharmaraja et al. 2014) proposed a model to protect the IMS network from degradation due to messaging overload via resource reservation planning, the throughput values of SIP signalling traffic was improved for both INVITE and non-INVITE SIP messages type.

3.7 IMS PERFORMANCE BENCHMARK AND PERFORMANCE METRICS

Due to the large number of technological variables, there is a need for putting in place a reasonable ground rules to be able to define a common architecture. Standardisation will help both Service Providers (SPs) and vendors to work independently via agreeing on a common ground and to be able to come up with products, standards and systems that comply with the technology guidelines and seeked service quality. A performance benchmark of IMS was introduced by the European Telecommunication Standards Institute (ETSI) in its technical Specifications listed in (ETSI 2013, ETSI 2013, ETSI 2015, ETSI 2015).

In (ETSI 2013), an IMS benchmark was introduced for the objective of getting a better understanding of IMS systems and how to set the performance metrics of interest to evaluate the performance. The System Under Test (SUT) is presented with a workload, traffic generated by simulated UEs, by the test system. The traffic set and rate is defined to be able to control the test load and get statistically enough collected data. The call attempt, which is referred as a scenario attempt for a more general meaning, could reflect either registration, call session establishment, messaging or any other IMS related service type. The result of the scenario attempt may finish with success, fail, or out of range success that happens if it exceeds certain threshold set by the design objective of the scenario. The term “Inadequate Handled Scenario Attempts” is referred whenever there is a scenario has failed to finish or successfully finished but exceeded the threshold time. Figure 3.5 shows an example system performance based on the previous discussion. The scenario attempt rate (Scenario Attempts per Second) is made steady by the traffic-time profile for certain period of time. In figure 3.5, for example, the rate of attempts has been increased during the first 10 minutes till it ramped up to 120 attempts per second and remained relatively constant for 30 minutes. Then the rate was increased again after 40 minutes of the beginning of the test to get another set of results. The results shows the variation in the percentage of inadequate handled scenarios frequency and the successful attempts rate after being averaged over the steady-state phase excluding the transient state spikes.

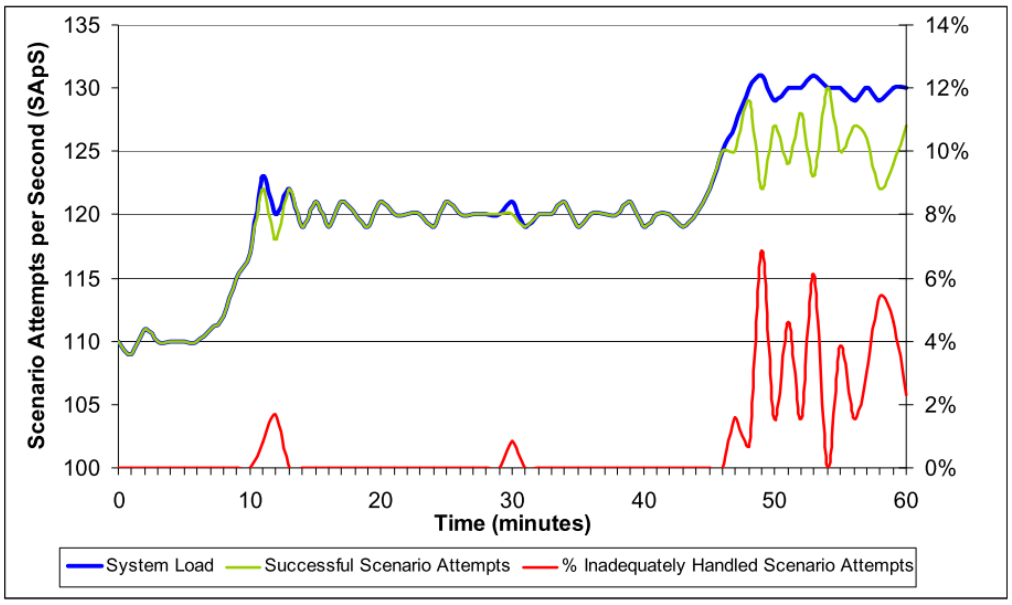


Figure 3.5 Example System Performance (ETSI,2015)

The IMS benchmark information model can be subdivided into three main entities; the Use-case, Benchmark Test, and Test Report. The Use-Case defines the call flow, the load profile, Metrics, Use-Case outcomes, scenarios/Scenario Attempts, and Design Objective. The Benchmark test measures the behaviour of a user population by defining the traffic set, the background load, Traffic-time Profile, and Test Parameters. It takes into account the collection of different types of traffic types which is referred as Traffic Set. The background load is presented to the system under test for the sake of replicating real-world scenarios. The Traffic time profiles describes the arrival rate as a function of time during the execution of the test. The test parameters should be specified by the benchmark to control the behaviour of the test script. Table 3.2 list example data elements used to configure the whole system. Finally the Benchmark Report is the collected data that could be represented in terms of charts to reflect the overall system performance and behaviour over certain elapsed time of the test. A report of relevant performance metrics are used to estimate the benchmark results and compare it among different other benchmarks result of different systems under test.

Performance Metrics are collected either in real-time or after the execution of the test. Sample Benchmark Metrics are listed in table 3.3. The parameters are either related to the IMS/SIP signalling or the speech quality parameters analysis.

Table 3.2 Test Parameters

Parameter	Description
Start time	Amount of time that a system load is presented to a SUT at the start of the test
Stop time	Amount of time that a system load is presented to a SUT at the end of the test
Total Provisioned Subscribers	The Number of simulated subscribers provisioned in the network
Percent Simulated Subscribers	The average percentage of simulated subscribers
Simulated Maximum Simultaneous call legs	The number of simulated maximum simultaneous call legs
Traffic per Subscriber	Traffic per subscriber, default value 0.1 Erlang
PX_Percent Registerd Subscribers	The average percentage of simulated subscribers that are registered simultaneously
PX_Percent Roaming Subscribers	The average percentage of simulated subscribers that are roaming (i.e. register in a no-local network)
PX_simulated Subscriber Registration Parameters	Parameters and Distributions of the probabilistic model simulated subscriber operation
MHT	Mean Holding Time of a call (default 110 seconds)
Ringing Time	Duration between (180 ringing and 200 OK INVITE) default value 1s to 5 s
NoS	number of subscribers originating traffic per subscriber
CAPS/BHCA	Call attempts per second/busy hour call attempts
WLF for Call Controller	The workload factor for Call Controller for specific configuration. Default value 1 to 3
WLF for Gateway Controller	The workload factor for Gateway Controller for specific configuration. Default value 1 to 3
WLF for MGW	The workload factor for Media Gateways for specific configuration. Default value 1 to 3
TDM Trunks	Number of TDM trunks
ETH	Number of ETH Connections
Type of call	MMTel to MMTel
	Video Telephony
	IMS-PES to IMS-PES
	MMTel to IMS-PES
Protocol call type and interfaces	SIP-I, ISUP, SIGTRAN (M2PA; M2UA; M3UA; SUA) SIP II NNI (Ici,Izi) , SIP NNI (Mx Interface)
MGCF/MGW/I-BCF/TrMGW performance tests	SIP-I to SIP-I , SIP-I to ISUP , SIP-I to NNI, NNI to NNI , SIGTRAN to SIGTRAN (M2PA; M2UA; M3UA; SUA) , ISUP (SIGTRAN) to NNI
Transport Interfaces	Voice over LTE (VoLTE) (LTE-Uu, S1-U, S-11, S6a, S11, S5/S8, Rx, Gx, Mw, ISC and Ut interfaces)
	Ethernet
	xDSL
	ISDN
	POTS (Z)

Table 3.3 Benchmark metrics examples

Delay parameters	**Description**
Call request delay	Call request delay is defined as the interval from the instant at which the INVITE message has been received from the SIP subscriber until the 100 Trying from the SBC/P-CSCF is passed back to the subscriber.
Alerting sending	Call request delay is defined as the interval from the instant at which the 180 Ringing is received from the terminating subscriber until the 180 Ringing is passed back to the originating subscriber.
Call set up delay	The time interval starts when the digit collection function determines that the address information received in the INFO or subsequent INVITE message is sufficient for session initiation, and ends when the INVITE message on the Ic or terminating Gm interface has been sent.
Through-connection delay	The through connection delay is defined as the interval from the instant that the 200 OK message is received from the called user at the terminating Gm interface until the through connection is established and available for carrying traffic and the 200 OK message has been sent to the calling user on the originating Gm interface.
Connection release delay	The through connection delay is defined as the interval from the instant that the 200 OK message is received from the called user at the terminating Gm interface until the through connection is established and available for carrying traffic and the 200 OK message has been sent to the calling user on the originating Gm interface.
Connection release delay	Connection release delay is defined as the interval from the instant when a BYE message is received at the originating or terminating Gm interface until the instant when 200 OK is sent and a corresponding BYE message is sent at the terminating or originating Gm interface respectively.
Speech quality analysis	
Speech Quality	PESQ (Recommendation ITU-T P.862) and Recommendation ITU-T P.862.1.
	P.863 POLQA "Perceptual objective listening quality assessment" (Recommendation ITU-T P.863)
Speech Level - Active Level	Recommendation ITU-T P.56 .
Speech Level - Peak	
Speech Level - Noise	
Speech Level - Signal to Interval Noise	

3.7.1 Subsystem Configuration and Benchmarks

In (ETSI 2013) ETSI has described the benchmark a for the IMS based services. In (ETSI 2013) ETSI has defined the general core concepts and terminologies related to IMS benchmarking, a more specific scenarios and use-cases benchmarking in addition to more detailed scenario specific performance metrics and a set of parameters configured for the system under test is described in (ETSI 2013). The goal is to define the main functional elements and to configure it whether it was implemented in hardware or software, the QoS specification measurements need to be defined at each interface connected the system under test to other entities in the system. The hardware-specific measurements such as CPU and memory utilisation should be collected during the system under test after being configured. Finally, the system under test interfaces should be well defined in case it was decided to take into account the performance implications over the overall system performance. Transport Layer Security (TLS), Internet Protocol Security (IPsec), Datagram Transport Layer Security (DTLS) are examples of such interfaces.

3.7.2 Traffic Sets and Traffic Profiles

In (ETSI 2015), ETSI has defined traffic set, benchmark testing procedure, and traffic-time profiles for benchmarking. The benchmark will help system designer and manufacturers in comparing the performance of two products or designs or for predicting the performance. In (ETSI 2015), the test scenarios are defined as protocol diagram without dictating the specific implementations of the test scenarios, the scenarios are either implemented by a commercial test system provider or by benchmark test implemented in hardware or software or a combination of both.

The traffic set is normally composed of mixture of test scenarios that is generally governed by the relative frequency of occurrence and the traffic-time profile. The occurrence frequency is specified by traffic set parameters, whereas the traffic-time profile is part of the testing system and it controls the arrival rate of scenarios and how it changes with time. Table 3.4 and table 3.4 shows sample traffic sets and traffic-time profile respectively.

Table 3.4 Sample Traffic Sets

Use Case Section	Test Scenario	Scenario Arrival Distribution	Scenario Duration Distribution
Registration/ De-registration Use-Case	Successful initial registration with unprotected REGISTER requests on the SIP default port values as specified in IETF RFC 3261	N.A	N.A
VoLTE to ISDN	Basic call. The call is released from the calling user This scenario represents the case when the call establishment is performed. The call is released from the calling user. Ensure that in the active call state the voice transfer is performed	Poisson, mean selected by traffic-time profile	Exponential, mean 120 s
VoLTE to ISDN	Basic call The call is released from the called user This scenario represents the case when the call establishment is performed. The call is released from the called user. Ensure that in the active call state the voice transfer is performed	Poisson, mean selected by traffic-time profile	Exponential, mean 120 s
VoLTE to PSTN	Basic call. The call is released from the called user. This scenario represents the case when the call establishment is performed. The call is released from the called user. Ensure that in the active call state the voice transfer is performed	Poisson, mean selected by traffic-time profile	Exponential, mean 120 s
VoLTE to PSTN	Basic call. The call is released from the calling user This scenario represents the case when the call establishment is performed. The call is released from the calling user. Ensure that in the active call state the voice transfer is performed	Poisson, mean selected by traffic-time profile	Exponential, mean 120 s

Table 3.5 Traffic Time Profiles

Traffic-time Profile Parameter	Traffic-time Profile Value
PX_SimultaneousScenarios (SIMS)	2
TotalProvisionedSubscribers	100 000 Subs
SApSIncreaseAmount	10 SApS
StepTime	30 minutes
StepNumber	3 steps

3.7.3 Reference Load Network Quality Parameters

In (ETSI 2015), ETSI has defined a reference load definition and values that can be used by system designers as benchmark. There are two reference loads; the first one is defined for ISDN/PSTN based systems, and the second one is for the VoLTE/IMS based systems. The reference load definition for ISDN/PSTN is described in ITU-T Q.543 , and the reference load definition for VoLTE/IMS is described in ETSI TS 101 563 , the values for ISDN/SIP interworking is described in ETSI TS 183 036 and the PES procedures is based on the IMS/PES emulation specified in ETSI TS 183 043 . Finally all SIP and Session Description Protocol (SDP) is based on ETSI TS 124 229 .

Table 3.6 shows sample performance metrics for the internal traffic within the VoLTE/IMS or ISDN/PSTN networking interactions. The delay values are to be referenced by system designers and vendors to ensure that the system entities and interfaces comply with the QoS requirements.

Table 3.6 Sample VoLTE/IMS Performance Metrics

Meaning of timers	**IMS, PES equivalent**	**Reference Load A**		**Reference Load B**	
		Mean Value	**95 % probability of not exceeding**	**Mean Value**	**95 % probability of not exceeding**
IMS SUBSCRIBER Local exchange call request delay.	IMS Call request delay is defined as the interval from the instant at which the INVITE message has been received from the SIP subscriber until the 100 Trying from the SBC/P-CSCF is passed back to the subscriber.	≤15 ms	≤20 ms	≤30 ms	≤40 ms
IMS SUBSCRIBER LINES 180 sending Delay for Internal traffic.	IMS For calls terminating sending delay is defined as the interval from the instant that a 180 message at the Gm interface has received and 180 is sent on the Gm towards the calling subscriber.	≤100 ms	≤150 ms	≤200 ms	≤250 ms
VoLTE SUBSCRIBER LINES 180 sending Delay for Internal traffic.	VoLTE For calls terminating sending delay is defined as the interval from the instant that a 180 message at the VoLTE - UE interface has received and 180 is sent on the VoLTE - UE towards the calling subscriber.	≤150 ms	≤200 ms	≤250 ms	≤300 ms
IMS SUBSCRIBER Call set up delay using for Internal traffic.	IMS Session initiation call set-up delay is defined as the interval from the instant when the INVITE signalling information is received from the calling user on the originating Gm interface until the instant when the corresponding INVITE signalling information is passed on the terminating Gm interface to the called user.	≤250 ms	≤350 ms	≤450 ms	≤550 ms
VoLTE	VoLTE Session initiation call set-up delay is defined as the interval from the instant when the INVITE signalling information is received from the calling user on the originating VoLTE - UE (ECM idle) interface until the instant when the corresponding INVITE signalling information is passed on the terminating VoLTE - UE (ECM Idle) interface to the called user with QCI 1 (see note 2).	≤1800 ms	≤1900 ms	≤2000 ms	≤2100 ms
VoLTE	VoLTE Session initiation call set-up delay is defined as the interval from the instant when the INVITE signalling information is received from the calling user on the originating VoLTE - UE (ECM Connected) interface until the instant when the corresponding INVITE signalling information is passed on the terminating VoLTE - UE (ECM Connected) interface to the called user (see note 3).	≤280 ms	≤380 ms	≤500 ms	≤ 600 ms
IMS Through-connection delay for Internal traffic.	IMS The through connection delay is defined as the interval from the instant that the 200 OK message is received from the called user at the terminating Gm interface until the through connection is established and available for carrying traffic and the 200 OK message has been sent to the calling user on the originating Gm interface.	≤100 ms	≤150 ms	≤200 ms	≤ 250 ms
VoLTE	VoLTE The through connection delay is defined as the interval from the instant that the 200 OK message is received from the called user at the terminating VoLTE - UE interface until the through connection is established and available for carrying traffic and the 200 OK message has been sent to the calling user on the originating VoLTE UE interface.	≤ 150 ms	≤ 200 ms	≤ 250 ms	≤ 300 ms
IMS SUBSCRIBER Connection call release delay for Internal traffic.	IMS Connection release delay is defined as the interval from the instant when a BYE message is received at the originating or terminating Gm interface until the instant when 200OK is sent and a corresponding BYE message is sent at the terminating or originating Gm interface respectively.	≤100 ms	≤150 ms	≤200 ms	≤250 ms
VoLTE - IMS SUBSCRIBER Connection call release delay for Internal traffic.	VoLTE Connection release delay is defined as the interval from the instant when a BYE message is received at the originating or terminating VoLTE - UE interface until the instant when 200OK is sent and a corresponding BYE message is sent at the terminating or originating VoLTE - UE interface respectively.	≤150 ms	≤200 ms	≤250 ms	≤300 ms

3.8 IP MULTIMEDIA SUBSYSTEMS QUEUING MODELS

In (ZHU 2003), the authors has introduced delay analysis model of IMS. The study tries to come up with a capacity estimate of the system along with the average time the task remains in the system before being completed. Using M/M/1 queue model (with Poisson distribution inter-arrival time and exponential distribution serving time) for the delay analysis model calculations,

The utilization for each entity is calculated by multiplying the job arrival rate by number of hits (call setup signalling hits) for each server, then multiplying by the average service time at each server. And finally to get the overall call setup delay we add all the calculated delays for each server or entity found in the signalling path.

Similarly, the average delay for M/D/1 model was calculated based on the assumption that all servers have the same service time ($\bar{x}$), which makes the average waiting time (W) for all servers the same and as given by equation (3.16).

in (RAJAGOPAL 2006), the author proposed a service models for IMS system, the service model, which depend on the formulation of queuing models and SIP traffic load characterization, helps in predicting the performance trend of the network and act as a control mechanism to optimize the network properties and performance. Using M/G/1/∞ model for three feed-forward tandem servers connected sequentially (P-CSCF, I-CSCF, and S-CSCF). The average queue waiting and service time for the three servers can be calculated using this model.

The average time represent both the queue waiting time and the average service time for one single server modelled as M/G/1 queuing system. It is assumed that the total session establishment time τ is more than or equal to the total average delay for K tandem servers in the network, this to prevent queue overflow condition. The total average delay experienced for K servers will be simply ($K . T_{avg}$).

Similarly, in (RAJAGOPAL 2006), for the same sequential feed-forward servers model, each server was assumed to have an M/M/1/∞ queue model (with exponentially distributed service time and infinite queue length).

Assuming the number of servers in the system and the number of servers in the call setup signalling path are the same and denoted as (K). Assuming as well that the servers in the network are independent from each other and each has the same service probability distribution function.

3.9 STANDARDS AND TECHNICAL REPORT FOR MCS

Standardisation bodies, such as 3GPP, has shown interest in developing a platform for the mission critical systems and integrating it via set of standard reports. This section will briefly

discuss the updates in the field related to be achieved so far regarding the standardisation efforts of the MCS.

It was in Rel-13 3GPP has enabled MC services in LTE via introducing multicast bearers in LTE following the standardisation of Group Communication System Enablers (GCSE) and Device to Device (D2D) communications for both the group communication and the Direct Mode of Operation (DMO) consequently. This opened the door for more discussions whether future mission critical systems should be implemented over existing LTE network, separate LTE network or a combination of both (3GPP TS 24.484,2017). The three types of mission critical systems has been discussed and introduced in Chapter 1.

In Rel-13 3GPP standardised Mission Critical Push to Talk (MCPTT) by the end of 2016 which is mainly an application layer standard that support the PTT operation over the existing LTE standard. MCPTT was a major step toward fulfilling the market needs for a functional MCS. In Rel-14 3GPP added in 2017 additional MC enhancements and services to its already existing MCPTT application standard (3GPP TS 23.379,2017), such as; enhanced MCPTT, Mission Critical Data (MCData), Mission Critical Video (Video) (3GPP TS 22.281,2018), and a general framework that facilitates the mission critical services (3GPP TS 23.379,2017). Such efforts required a new set of protocols and new security enhancements to the already existed MCPTT application in Rel-13 (3GPP TS 33.180, 2017).

In Rel-15 the MCS is further improved in aspects related to interconnection between 3GPP defined MC systems, Interworking between 3GPP defined MC system and legacy systems such as TETRA or P25 for voice and short data services (3GPP TS 23.283, 2018), the MC service requirements for railway industry, and the mission critical service requirements for maritime industries.

3.10 CURRENT CHALLENGING ISSUES

This section classify the challenges that was covered but with potentials of further improvement.

3.10.1 Highlights and Classification of Current Research

In this section, the studied literature will be classified into categories that reflect the advancements in different layers within the ISO/OSI Networking model. Then an evaluation of the missing part in each category and its implication over the proposed solution related to the mission critical aspects will be presented.

Based on the literature and related studies presented in the previous section. It is found that there is a need for enhancing the performance of the communication systems used for mission critical applications for a more scalable and reliable approach compared to the currently deployed ones. Based on (Paper 2016), Nokia has proposed its own vision for the future Mission Critical Communication Systems. Less end-to-end latency and more reliable communication system are the core requirements of any proposed solution has the potential to be the winning technology candidate considered for the future Mission Critical Communication System as part of the 5G technology standard. Other requirements for the future MCS under 5G communications umbrella include as well having the ability to adapt more diverse

multimedia applications with higher data rates and less latency in addition to the aforementioned higher reliable operating structure.

From the literature and related studied presented earlier in this chapter, it was found that most of the studies within the recent decade focused on improving the SIP or IMS performance. The studies can be classified into different general categories based on its contribution as shown in the following subsections.

3.10.1.1 Hosting SIP/IMS machine performance enhancement and evaluation

Many studies deployed a certain experimental setup and evaluated its performance with suggestion on how to improve it. In (Krishnamurthy and Rouskas 2016) (Krishnamurthy and Rouskas 2015) (Femminella, Maccherani et al. 2011) (Jia, Liang et al. 2008) (Wright, Nahum et al. 2010) has contributed in giving proposals for improving the hardware and/or the software running the SIP server machine. In (Krishnamurthy and Rouskas 2016) (Jia, Liang et al. 2008) and (Wright, Nahum et al. 2010) , the effect of increasing the number of CPUs over SIP server performance was investigated. The Operating System kernel process scheduler performance enhancement and its effect over the overall SIP performance was presented in (Krishnamurthy and Rouskas 2015). The influence of Java-based SIP or IMS application server performance issues over SIP/IMS overall performance was presented in (Femminella, Maccherani et al. 2011).

3.10.1.2 Performance evaluation of Transport Layer and other optional services protocols

In other research papers and articles, authors have given more attention for the transport layer, IPv6, or security layer performance issues and its implication on the overall performance of SIP/IMS performance. In (Dacosta, Balasubramaniyan et al. 2011, Hossain, Ariffin et al. 2011, Desai, Alagesan et al. 2012, Kulin, Kazaz et al. 2012, Shen, Nahum et al. 2012, Alshamrani, Cruickshank et al. 2013, Baghdadi and Azhari 2013), researchers has investigated the performance improvement or the overhead introduce due to the application of optional SIP services such as the TLS, IPv6, and IPsec. In (Hossain, Ariffin et al. 2011, Alshamrani, Cruickshank et al. 2013) the SIP performance using IPv6 overhead was evaluated for Mobile Ad hoc Networks (MANETs) and indoor environments respectively. In (Happenhofer and Reichl 2010) the quality of SIP signalling was evaluated over different transport protocols (UDP and TCP). In (Baghdadi and Azhari 2013) the capabilities of connection oriented transport protocol, such as TCP, was investigated to enhance SIP performance. In (Dacosta, Balasubramaniyan et al. 2011, Desai, Alagesan et al. 2012, Kulin, Kazaz et al. 2012, Shen, Nahum et al. 2012), security aspects in terms of improving the security or studying the overhead of applying security measure for SIP was evaluated.

3.10.1.3 Databases and DNS lookup performance implications

In (Chen, Cheng et al. 2014, Kellovsky and Baronak 2014), the performance of the database and its signalling overhead was evaluated for both SIP servers and for the more complicated IMS infrastructure. Both SIP server registrar and IMS network has a database, named as Home

Subscriber Station (HSS), of registered users along with user status and set of supported user-level services and network-level capabilities that are shared with other servers.

In (Kist and Harris 2003), the authors has evaluated the delays due to SIP signalling in 3GPP IMS networks. Possible sources of delays and its influence over the QoS is discussed. DNS system delay was among the delay components studied which could be of great impact especially in large scale IP networks. The study shows that DNS request retransmissions impose significant delays that may affect the call setup time.

3.10.1.4 SIP service modelling and performance evaluation using Queuing models

Performance modelling via queuing models analysis of SIP/IMS signalling performance was discussed in (Gurbani, Jagadeesan et al. 2005, Subramanian and Dutta 2008, Subramanian and Dutta 2009, Subramanian and Dutta 2009, Subramanian and Dutta 2010, Krishnamurthy and Rouskas 2013) and in (ZHU 2003, RAJAGOPAL 2006). The models can be considered as a preliminary performance evaluation of real networking implementations. Queuing models are helpful for bottle-neck analysis related studies especially in subsystems of multi-server structure such as IMS. Most of the queuing models presented relies on model parameters that replicate real network behavioural statistics. Table 3.6 and table 3.5 in the previous section presented sample parameters of such models.

3.10.1.5 Capacity and Scalability analysis with load balancing performance evaluation

Scalability analysis with load balancing performance evaluation studies was investigated in (Montagna and Pignolo 2008, Subramanian and Dutta 2009, Jiang, Iyengar et al. 2012, Al-Doski, Ghimire et al. 2013, Mishra, Dharmaraja et al. 2014). In (Subramanian and Dutta 2009), the multithreaded property of the kernel was modelled and results showed that the SIP server can meet the scalability requirements using multithreaded processing techniques. In (Mishra, Dharmaraja et al. 2014), a scheme to enhance the SIP performance under overload state via proper resource reservation planning and logical separation between INVITE and non-INVITE messages. in (Montagna and Pignolo 2008), two algorithms to control the SIP server overload were considered, the first one controls threshold values based on INVITE message arrival rate, and the second algorithm controls the overload via monitoring the CPU load. In (Jiang, Iyengar et al. 2012), another load balancing algorithms is proposed for SIP server clusters, the performance of the different algorithms was compared among each other. In (Al-Doski, Ghimire et al. 2013), discussed in general the scalability of IMS networks with based on different users distribution models and SIP registration generation rates and it is integrity with other access technologies. In (Baghdadi and Azhari 2013), an approach that tries overcoming the SIP overload impact over its performance using the connection oriented transport layer capabilities was proposed.

3.10.1.6 Cross layer optimization studies

Most of the studies in the literature are tightly coupled with certain access technology and transport network. SIP as a signalling protocol is part of the application layer domain and embedded in the network layer generated packets. In the Next Generation Networks (NGN)

standards, the service related functions and protocols was separated from the underlying access technology functions and specification for the sake of providing seamless connectivity. This impose better integration among different access technologies connected to the same core network. Vendors will be able to have the flexibility of providing the needed products and services by following the standard interface guidelines connecting two-layer domains. Signalling complexities will be reduced if the network layer (IP layer) is deployed at each entity between end users.

Operating in all flat IP architecture access technologies, such as LTE, introduces cross optimization challenges at different layer of the ISO/OSI model. Despite the many advantages of isolating the access layer services from the higher layers abstracts, many researchers has limited their efforts in one abstract domain without taking into account the chances of performance enhancement via cross layer optimizations techniques.

In many research papers mentioned in the literature study, it was obvious that the enhancements was strictly limited in one layer abstract without taking into the account the performance evaluation through cross layer optimization methods. In other research papers, the underlying access technology influence over SIP QoS was investigated to present a more integrated end-to-end analysis. In (Kueh, Tafazolli et al. 2005), SIP performance was analysed over satellite UMTS network, the access technology effect over SIP end-to-end performance is taken into account. In (Baudoin, Gineste et al. 2015), authors propose a QoS framework for satellite communication that is derived from the terrestrial IMS solutions to provide end-to-end QoS support.

3.10.2 Further improvement Highlights

According to the literature presented in the previous sections, we can list the missing, or not investigated enough, research items that are related to providing generic MCS with IMS based solution in the core network as follows:

1) Many proposed solutions to improve the hosting SIP/IMS machine was configured and validated for certain experiment setup that evaluate local parameters without taking into the account its influence over the end-to-end performance. The performance of specific hardware and/or software design is crucial in any evaluation method, but it would not be meaningful without reflecting the performance improvement and evaluation over the general and wider domain contribution approach.
2) Many studies has studied the main core IMS entities (i.e. P-CSCF, I-CSCF, and S-CSCF) in depth. But it was noticed that few of them has deeply analysed the effect of database and DNS lookup performance over the entire system. Testbed experiments studies have an advantage over system modelling and simulations studies in terms of results accuracy.
3) The queuing models presented before was a good way to evaluate the system performance and to have a proper explanation of system responses and delays under different traffic loads. Having an accurate and close to real life scenarios model is of a great challenge, it was noticed that many of the aforementioned studied has either

oversimplified the proposed model, or omitted entire entities performance influence by assuming fixed performance variables values.

4) Many performance metrics are suggested by the literature or even standardized in many technical reports, table 3.7 shows sample performance metrics found in literature. However, it was noticed that single hop SIP/IMS or end-to-end SIP/IMS performance metrics and indicators are able to reflect the performance in specific domain or give a general description of the overall system performance without targeting the performance delays or defections in other layers or subsystems. Session initiation Delay (SID), for example, calculates all delay components of the whole networking layer all the subsystem that SIP INVITE signal along with its provisions travel back and forth toward its final destination. Although it is good to know the SID values to evaluate the overall system performance in which the service operators may use it to monitor their SLA level, it would be useless whenever there is a need for finding impairments and defections in certain interfaces connecting different entities in the middle of the signalling path. As presented before in the literature, the other single hop quality of SIP signalling metrics, which count delays and losses over single interface, is good enough for bottle-neck analysis and system performance evaluation, but it targets general networking traffic improvement without having enough resolution of the wider SIP end-to-end performance vision.

Table 3.7 Performance metrics in the literature

Performance Metric	Related to	References
Waiting time performance metric	Process Scheduler	(Krishnamurthy and Rouskas 2015)
Packet Drop Rate Performance metric	Number of CPUs	(Krishnamurthy and Rouskas 2016) (Krishnamurthy and Rouskas 2015)
Total Capacity	Capacity planning model	(Krishnamurthy and Rouskas 2016)
Service time performance metric	-System Hardware and Software -SIP server response time	(Krishnamurthy and Rouskas 2013) (Gurbani, Jagadeesan et al. 2005)
Average Delay and number of jobs, queue length	Queuing Model	(Subramanian and Dutta 2008, Subramanian and Dutta 2009, Subramanian and Dutta 2009, Krishnamurthy and Rouskas 2013) (Gurbani, Jagadeesan et al. 2005)
Server based System response	Java-based SIP Application Server	(Femminella, Maccherani et al. 2011)
Layer scalability	transport layer	(Baghdadi and Azhari 2013)
CPU utilization	Computer Architecture	(Subramanian and Dutta 2009)

Memory Utilization	Software and Hardware setup	(Subramanian and Dutta 2009)
average accepted Call Hold Time (CHT)	User profile model	(Subramanian and Dutta 2009)
Database query performance	HSS/Diameter domain performance	(Chen, Cheng et al. 2014)

5) Based on the reviewed literature, it was obvious to us that not enough research is targeting cross layer optimization techniques to boost the signalling and overall system performance. Most of the research was directed toward single subsystem or networking layer without looking at other layers/subsystems utilization opportunities. The MAC and physical layers are essential part of any communication system regardless of the access technology used. Different access technologies use its own version of Medium access control and physical layer signalling options. Other researches, presented before, shows that the access technology has great impact over application layer centric protocols such as SIP. For the future MCS applications and based on 5G systems design vision and goals, system delays should be minimised and overall system convergence and integrity should be maximised. This can be achieved by exploiting the physical layer capabilities and link it with the higher layers for best end user experience.
6) Based on 5G vision, scalable communication system is needed to be able to adapt the increasing number of users accessing the system. Increased coverage with more capacity is of great challenge and one of the hot research topics that is being investigated currently by researchers. Beamforming and Millimetre Waves (mmWave) channel models along with massive MIMOs deployment are one of the proposed candidate technologies for the future 5G physical layer deployments. Having Radio Access Network (RAN) technology that is supported by good enough Core Network (CN) able to serve end users within a wider coverage area and with enough bandwidth is of a great challenge. Device to Device (D2D), Machine to Machine (M2M), and Human to Machine (H2M) communication application along with the conventional human to human communications have introduced a wide range of applications and user demands that need a core network with convergence capabilities to be able to aggregate the traffic and monitor the QoS dynamically. Based on the literature presented before for both IMS and SIP signalling, we believe that there is a need for proposing a new framework design of IMS subsystem to be able to adapt the huge expected traffic in case of emergency situation to be considered as part of new MCS design within the future 5G technology standard.
7) In terms of scalability and load balancing studies, it was noticed that the literature missed detailed studies of the ability of different IMS entities to adapt huge IMS/SIP related traffic and providing simple though efficient load balancing techniques without dramatically altering the IMS structure design and interfaces.
8) In terms of the evaluation techniques presented in the literature studies. It was noticed that the authors has either used simulations, experimental testbed setups, or Queuing models to test their theories and validate the results. Some of them has made a comparison between the simulated and experimental results, others have compared between the modelled system expected results and the experimental results. All was carried separately and results were compared at the end. Based on the reviewed literature, still there is a chance to deploy new methodology to evaluate the theory and validate the result following a new approach that is based on mixed modelling and real time testing methods.

3.11 SUMMARY

The up to date related works and literature that is related to the SIP and IMS performance was presented in this chapter. The diverse enhancements in SIP and IMS performance was highlighted to get better understanding of the gaps in current research trends. Literature was classified into emerging trends of SIP and IMS performance and the intended research contribution was identified. According to the literature, the scalability was one of the major drawbacks of SIP servers and IMS subsystems.

It was shown that based on 5G vision, scalable communication system is needed to be able to adapt the increasing number of users accessing the system. there are many gaps in the field, for example, many proposed solutions to improve the hosting SIP/IMS machine was configured and validated for certain experiment setup that evaluate local parameters without taking into the account its influence over the end-to-end performance. It was obvious to us that not enough research is targeting cross layer optimization techniques to boost the signalling and overall system performance. Most of the research was directed toward single subsystem or networking layer without looking at other layers/subsystems utilization opportunities

Chapter 4: METHODOLOGY AND SIGNALLING ANALYSIS

4.1 INTRODUCTION

In this chapter, the framework design details will be presented and the system model parameters will be identified, the next chapters will refer to this chapter presented design methodology and framework for validation purposes. This chapter will introduce to the reader the general layered architectural model that is related to the scope of this study, signalling example to reflect the interfaces connecting different system component and entities between end-users at different layers, then the framework design of the proposed solution will be explained, the methodology will follow to evaluate the performance at different levels, and finally the algorithms and flowcharts to describe the evaluation logic.

4.2 RESEARCH METHODOLOGY

The methodology of the research followed in this research is based on the presented methodologies found in the literature and standard reports. IETF, for example, presented the methodology followed to evaluate SIP performance along with testing performance metrics in many technical draft reports such as in (D. Malas 2011) (Poretsky April 2015) (C. Davids November 12, 2014) to ensure that evaluation and benchmarking procedure carried out by research community is standardised and referenced to the standard. The recommendation and guidelines of IETF drafts will be followed throughout this research and will be listed in the following sections and chapter.

The logical methodology flow since the beginning of carried research can be summarised in the diagram shown in figure 4.1. Each step in the methodology followed will be explained in following subsections.

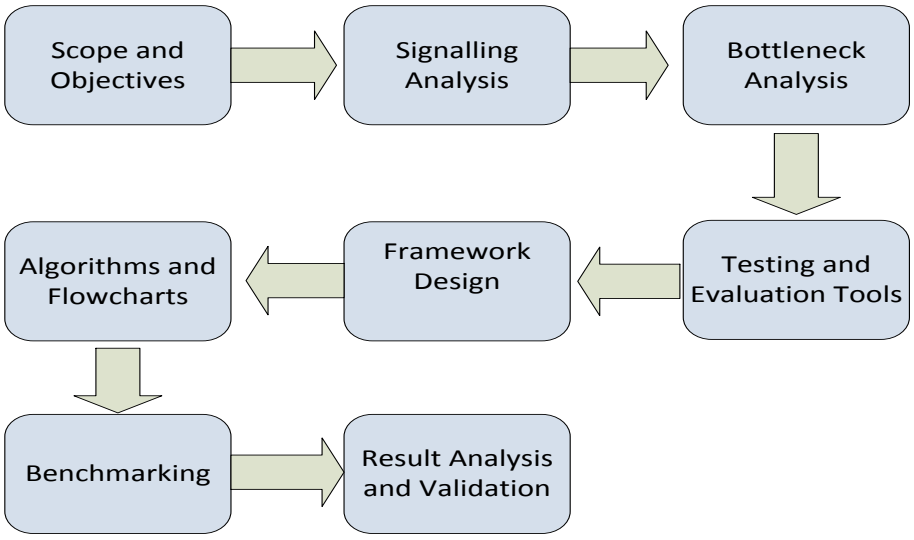


Figure 4.1 Research Methodology General Flow Chart

4.3 SCOPE AND OBJECTIVE METHODOLOGY

The IMS network is designed to provide a seamless connectivity for the underlying layers. Users use different access technologies that are connected to the same core network. The separation between the access plane and user plane is essential for service integrity especially in a vertical handover scenarios in heterogeneous communication systems, it also simplifies the design and operation of each plane via having set of interfaces for both data and control signals that ensure smooth transmission between different domains and plains. In figure 4.2 the general layered architecture model of a communication system with multiple access technologies is shown.

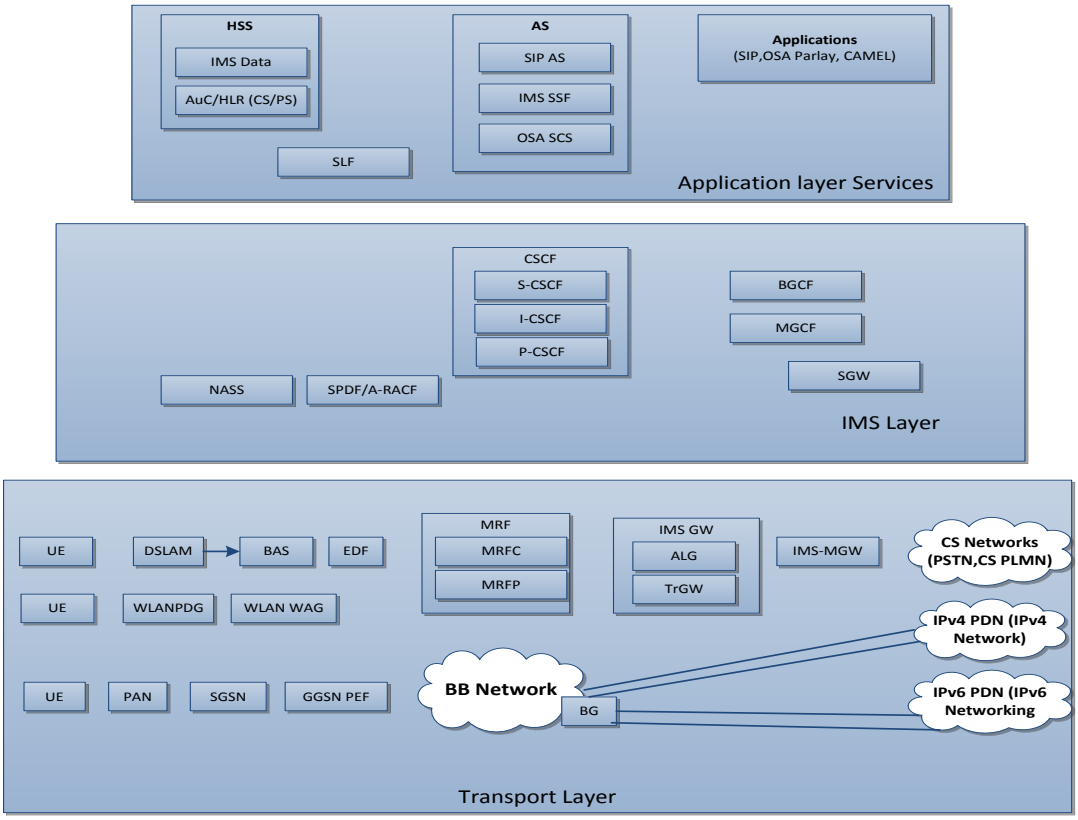


Figure 4.2 General Layer Architecture Model

As shown in the figure 4.2, different access technologies (LTE, fixed line, WLAN, broadband wireless WiMAX, etc.) are all connected to the Back Bone (BB) IPv4 or IPv6 based system through different data and control logical channel connections. The control plane is positioned logically above the user plane, the control plane controls the signalling of control information and media stream between different entities in the entire system. the application level (above the control and user level) deals with high level service management in which advance services are delivered seamlessly without worrying about the access and control layer control establishment and monitoring signalling complications. All the control signalling is aggregated by the control layer entities represented by IMS layer. The separation between different entities functionalities in the IMS layer was introduced as a unique feature for the NGNs designs compared to the previous generation networks, each entity is responsible for certain function

such as packet switch control, resource control, media control, service control and call control (Spirent 2013).

IP Multimedia Subsystems (IMS) (3GPP 2006) and Session Initiation Protocol (SIP) (Rosenberg 2002) performance play a major role in multimedia communication networks by altering the Key Performance Indicators (KPIs) related to the Quality of Experience (QoE) metrics of the end-to-end service. Registration Request Delay (RRD) is one of the SIP KPIs that also influence both IMS KPIs and end user QoE. Therefore, it is crucial to evaluate the performance of both SIP and IMS based on the RRD metric in order to give an indication of the overall system capacity and scalability potential.

The overall system capacity and scalability which is affected by added traffic introduced by more users who are trying to access the system provided services. This could happen in a mission critical communication system during natural disaster or large scale attack, where the system accessibility could be affected due to the sudden increase of number of users.

The need for a more detailed study of other SIP and IMS KPIs is vital to have a better understanding of the overall system performance which will enable us to take it a step further toward system performance enhancement and optimization to avoid single point of failure of the system.

A previously developed research methodology (Creswell 2009) was followed for both the qualitative and quantitative approaches when setting the parameters for all measurements and simulations. The methodology for deciding the qualitative values that need to be investigated can be summarized as follows:

1. Determine the challenges that need to be investigated within the scope of the study. While the research embeds several challenges, the focus was placed on the signalling domain especially between the end user and the core network and the signalling interface between the core network and IMS.
2. Determine the benchmark for what is considered acceptable SIP performance and decide on the metrics that will be measured and used to judge and compare the performance of setup.
3. Decide the appropriate simulation tools to generate the results from multiple sources that meet the appropriate comparison criteria based on the selected tool.
4. Determine the key factors that affect the SIP signalling, in addition to multimedia services operation in LTE and IMS that affect the overall QoS for the Mission Critical system.

The Quantitative Methodology to acquire the needed measurements is summarized as follows:

1. Develop a testbed for the IMS to determine the performance of the system. Then decide the performance metrics that need to be measured in order to facilitate comparison with other implementations and scenarios.
2. Develop a virtual machine to generate virtual clients along with IMS in order to facilitate comparison between the performance of the system with the real running testbed.
3. Develop a simulation project for both LTE and IMS over OPNET to benchmark their performance against the testbed implementation performance metrics.
4. Determine the variables, models, scenarios and parameters in OPNET that need to be adjusted and analysed to enable overall system's performance evaluation.

4.4 END-TO-END SIGNALING ANALYSIS

The methodology followed to determine gap in the field and potential improvement opportunities needs proper analysis of the signalling domain at different layers. For this purpose, in this section, the reader will get a better idea of the signalling challenges and the methodology followed to overcome the challenges. The signalling analysis is divided into three main categories; IMS related Signalling in the core network, Access Technology (i.e. LTE) related signalling, VoLTE Signalling. The following subsection will present more detailed view of the signalling analysis with the aim of load visualisation (collaboration diagrams) and sequence diagrams.

Following the layered architecture model presented in the previous section, the end-to-end signalling of a VoLTE call will be demonstrated and analysed in this section to be referred as a reference signalling model in the following sections and chapters. First, the end user side signalling perspective will be demonstrated to show the user interaction level in detail. Then, the detailed signalling accompanied by context diagrams of each subsystem will demonstrated.

4.4.1 The LTE Signaling from User Equipment perspective

In this section, signalling example of VoLTE signalling over IMS in LTE based systems will be demonstrated according to the UE perspective. There are two phases of signalling groups, the first one is the LTE system related signalling part as shown in figure 4.3 and the second one is mainly related to the IMS signalling part. Clearly the first signalling set initiated by the UE and ends with bearer service establishment before starting the IMS signalling part as shown in figure 4.3.

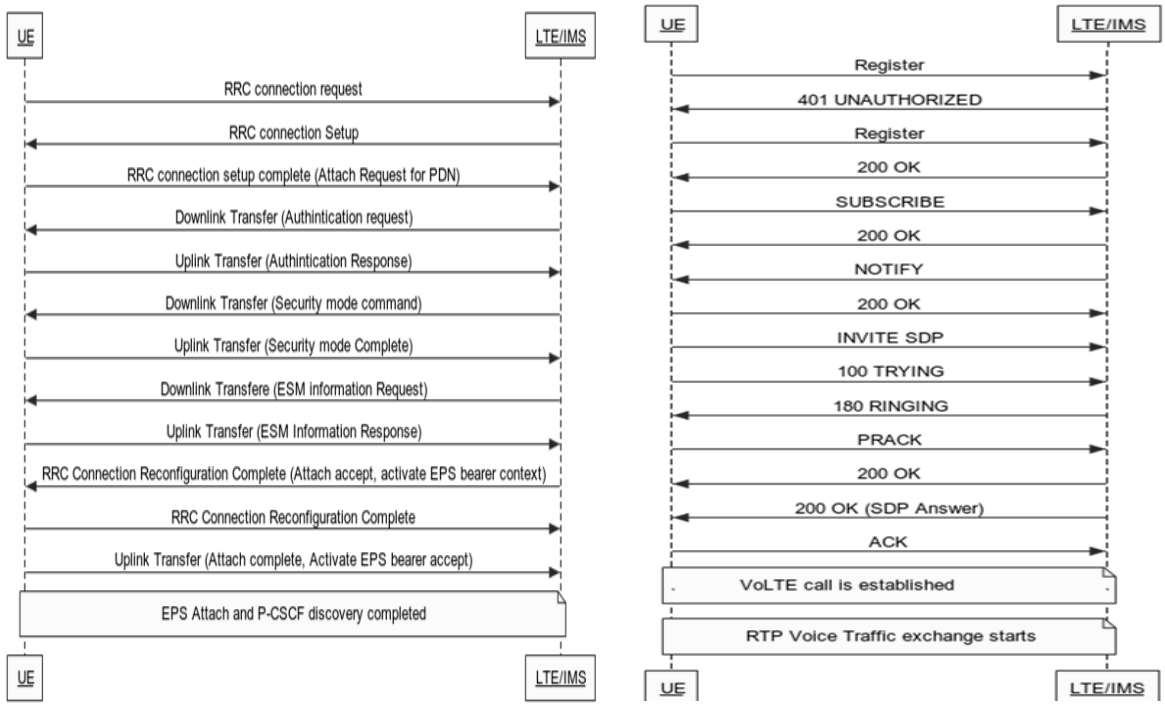


Figure 4.3 IMS signalling from user perspective

The UE first listen to the Master Information Blocks (MIBs) and System Information Blocks (SIBs), processing the information will make the UE able to reserve a temporary uplink resource slots to be able to initiate the remaining signalling sequence. The UE then starts the PDN connectivity process by sending Radio Resource Control (RRC) connection request which embeds the user identity information, the network will then reply with RRC connection setup message to indicate that a signalling bearer (to send forthcoming signalling messages) along with a Dedicated Control Chanel (DCCH) are established between UE and eNB, the UE finally acknowledge the reply with RRC connection setup complete message embedded with Attach request for Packet Data Network (PDN) connectivity, this message is forwarded later to the Mobile and Management Entity (MME) as an initial message in Non-Access Stratum (NAS) set of messages that maintain continuous communication with the user as it moves.

Following the attach request and the establishment of NAS signalling, the network starts user authentication process to check if the user is legitimate and allowed to use the network resources. Once the authentication challenge and response message exchange is finished, Data integrity is ensured via security mode command messages exchange which ensure encrypted signalling between UE and the network, finally the network protect the Evolved Packet System (EPS) Session Management (ESM) information, the network sends ESM information request and get back from the user the response describing the protocol configuration options.

Following the Authentication and security protocols exchange, the network sends RRC connection reconfiguration embedded with “attach accept” and “EPS bearer context activate” messages to allow the user to connect to additional radio bearers and to indicate that the Attach request to PDN is accepted. The UE replies back to acknowledge and complete the Attach process and accepts the activation of the EPS bearer. Now the UE is associated with PDN and able to send data along with control messages to the network. Finally the UE performs a P-CSCF discovery operation in which the IP address of the P-CSCF is resolved, normally the UE request it during the attach request process.

4.4.2 IMS Signaling Sequence

After being connected to a PDN bearer service, the UE is fully able to send and receive control and data messages. The UE establish Packet Data Protocol (PDP) context and then the IMS client in UE sends SIP Register request message to P-CSCF server. The SIP client registration process sequence diagram and context switch diagram are shown in Figure 4.4. Both diagrams are generated using Visual Studio programming based tool. Context diagrams will be used in this chapter to demonstrate the signalling flow and entity interactions.

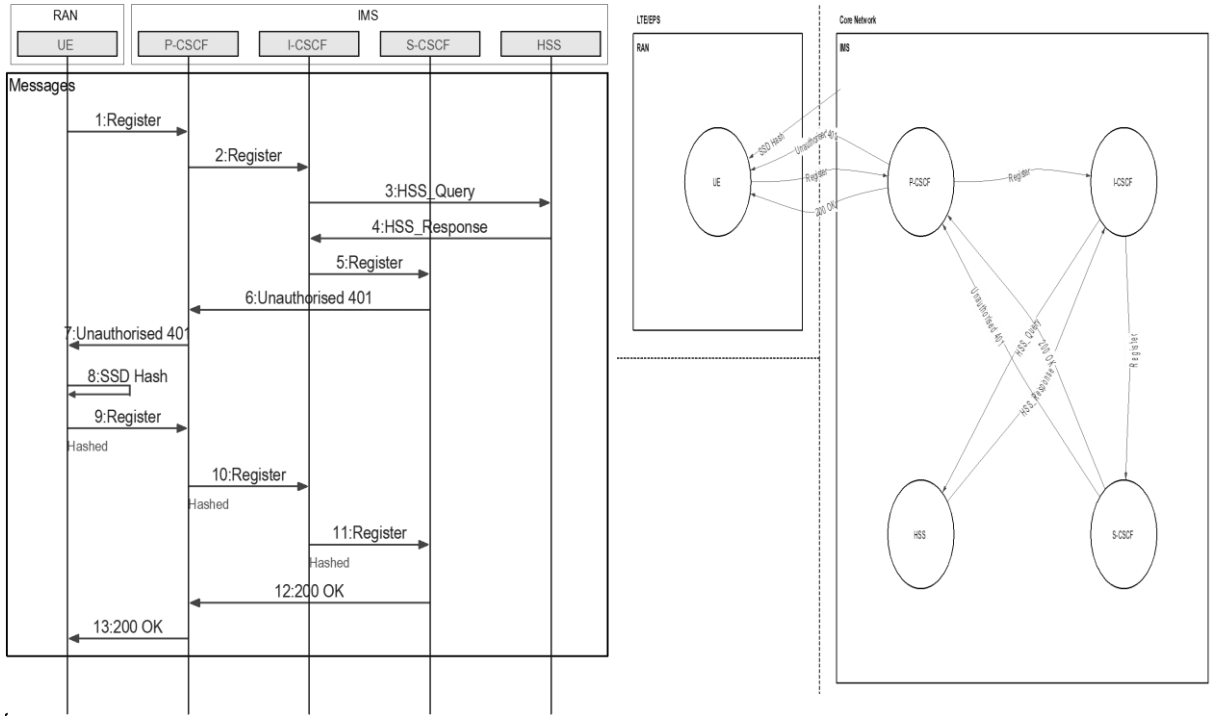


Figure 4.4 IMS Registration Sequence Diagram

4.4.3 LTE eNodeB Start-up and UE Setup Signalling

It is important to understand the access technology domain interactions level to better understand the delay and bottleneck possibility between the UE and the core network side. In this section the sequence signalling of S1 connection setup between the eNB and EPC, the MIB and SIB broadcast to the UEs, in addition to unicast of configuration information sent by eNB to specific UE will be demonstrated. All aforementioned procedures involves interactions between UE, eNB and MME. Figure 4.5 and 4.6 shows the sequence diagram of each interaction. The steps can be summarised as follows:

1) First, eNB initiates S1 connection with MME by sending S1 Application Protocol (S1AP) S1 setup request message, the MME replies back with S1AP S1 setup response message after accepting the request, now the S1 connectivity between eNB and EPC is established for forthcoming signalling.

2) eNB broadcasts MIB that carries the physical layer related information about LTE cell.

3) System Information Block Type 1 (SIB1) is broadcasted to all UEs, it contains information that is useful for the evaluation of UE ability to access the cell and sets the scheduling information for following system messages.

4) System Information Block Type 2 (SIB2) is then broadcasted to all cell UEs, it contains the radio resource configuration information common for all UEs.

5) System Information Block Type 3 (SIB3) is then broadcasted to all cell UEs, it contains intra frequency cell re-selection information that is common to all UEs.

6) System Information Block Type 4 (SIB4) broadcasted to all users loaded with information related to neighbouring cells intra frequency reselection process.

7) System Information Block Type 5 (SIB5) broadcasted to all users, it contains information related to inter frequency cell reselection.

8) System Information Block Type 6 (SIB6) broadcasted to all users, it contains intra frequency re selection process information.

9) System Information Block Type 7 (SIB7) broadcasted to all users with information related to inter-Radio Access Technology (RAT) cell re-selection.

10) System Information Block Type 8 (SIB8) broadcasted to all users with information related to intra-Radio Access Technology (RAT) cell re-selection.

11) System Information Block Type 8 (SIB8) broadcasted to all users with Home eNodeB name (HNB Name)

12) After the SIB broadcast messages, the UEs and eNodeB are ready for the random access procedure which is necessary for user synchronisation and reserving the needed resources in the uplink channel to be able to initiate RRC connection setup message.

13) The UE uses the uplink granted resource to send RRC connection request message over Up Link Control Chanel (UL-CCH).

14) eNodeB receives the request, sets up SRB1, and send back to the UE specific configuration via RRC connection setup message. Now a connection between UE and eNodeB is established.

15) The UE starts the attach and setup process of a default bearer.

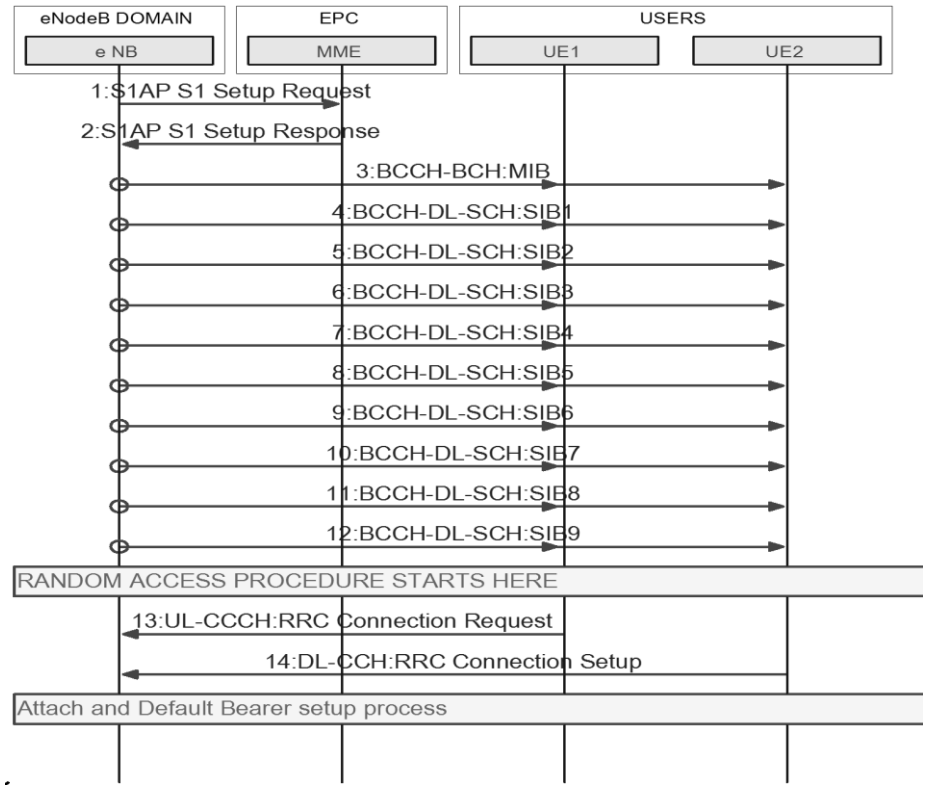


Figure 4.5 LTE eNB Start-up Sequence Diagram

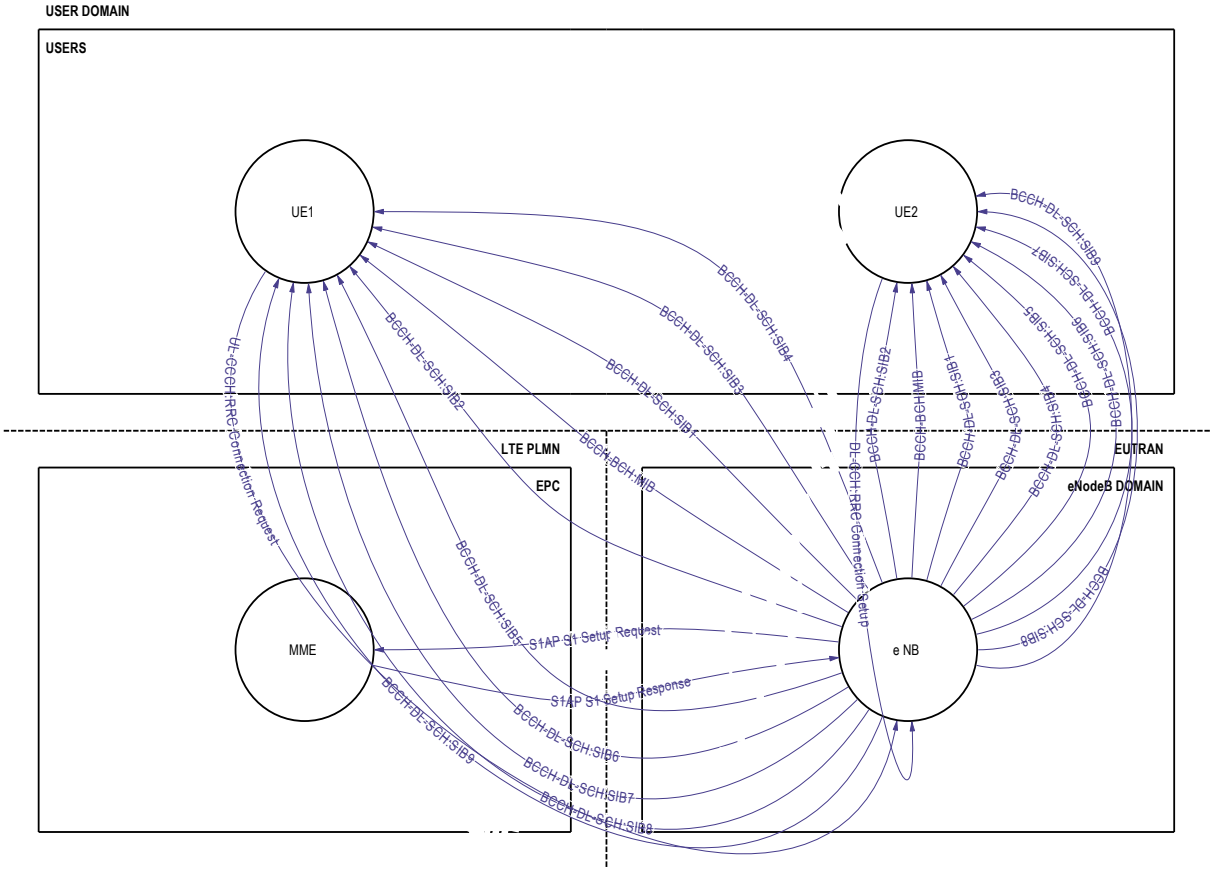


Figure 4.6 LTE eNB Start-up and Setup State Diagram

4.4.4 UE/eNodeB Random Access Procedure

As discussed in the previous subsection, following the exchange of MIB and SIB messages between eNodeB and UEs, the UEs starts the random access procedure. In this section, the signalling order and description of each step followed by signalling sequence and context switch diagrams will be demonstrated. The random access procedure is needed before starting the uplink data transfer process. Random access sequence diagrams and context switching diagrams are shown in figures 4.7, 4.8, and 4.9 respectively.

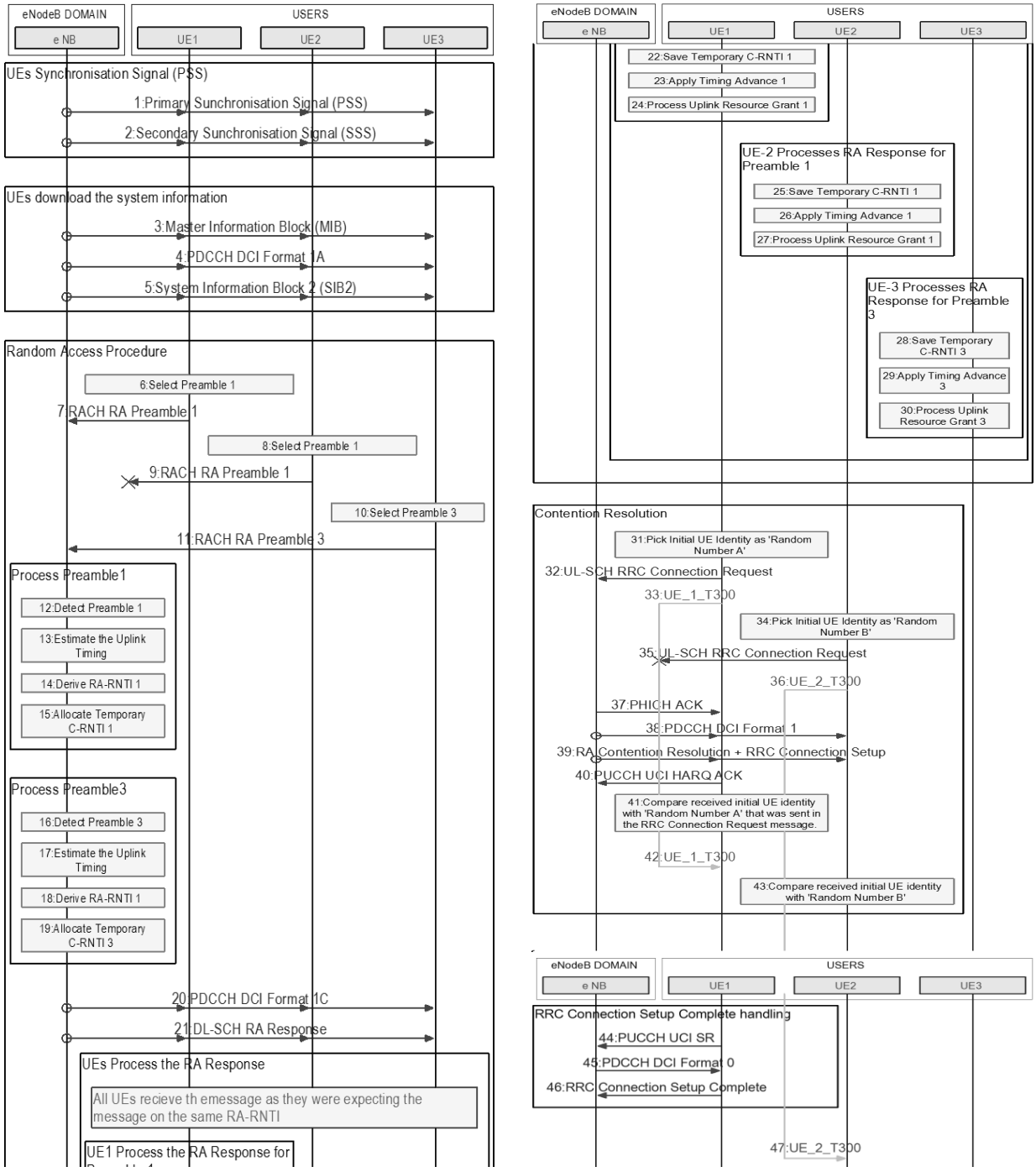


Figure 4.7 UE/eNB Random Access Procedure Sequence Diagram

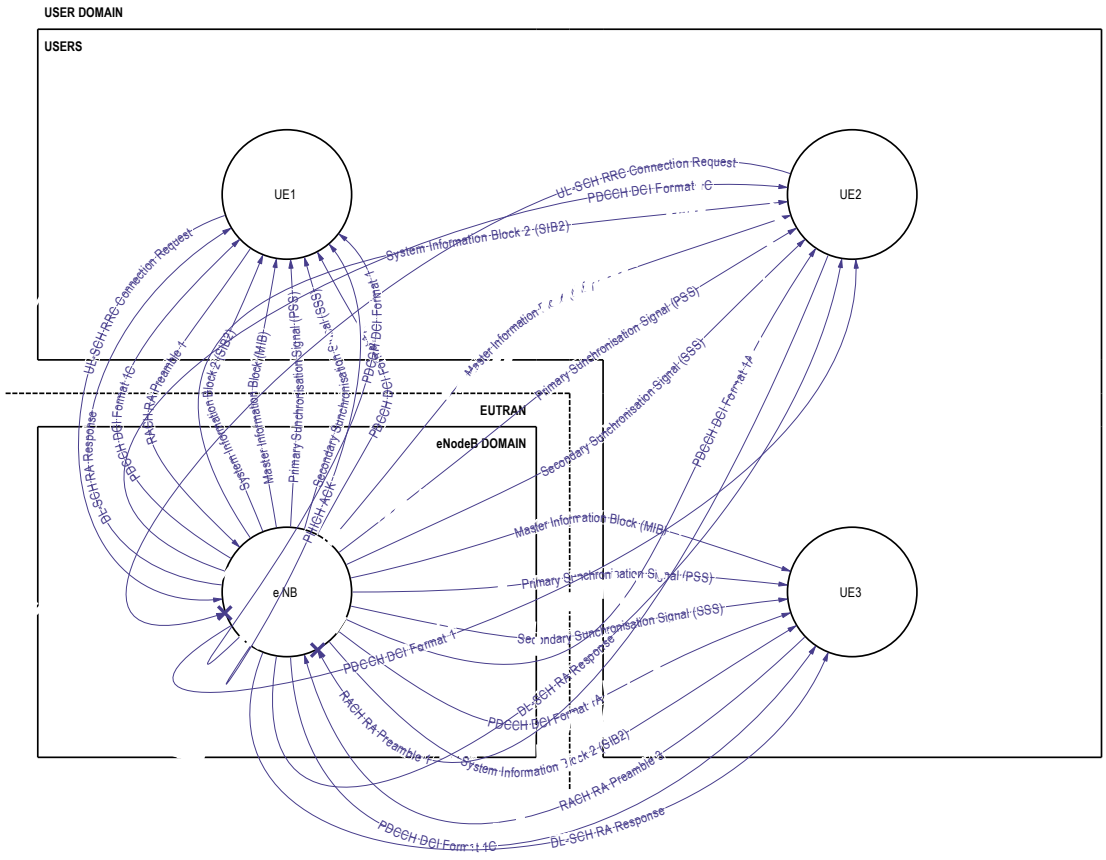


Figure 4.8 UE/eNB Random Access Procedure State Diagram

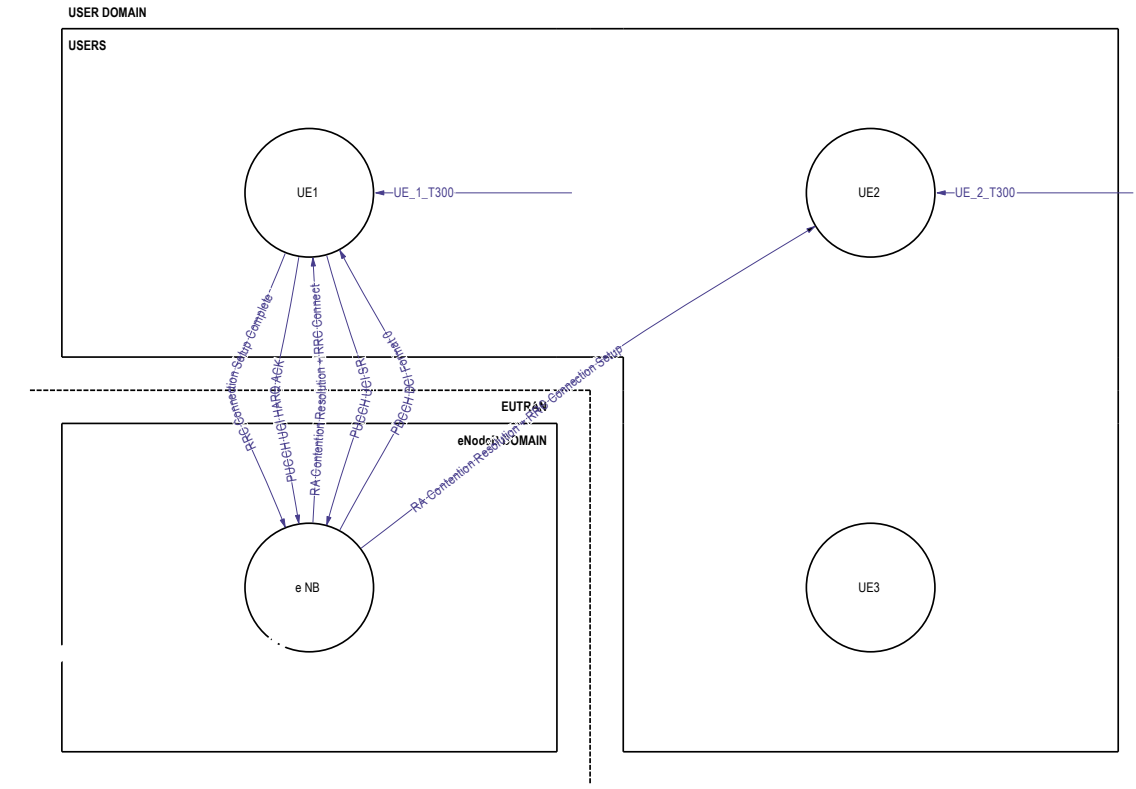


Figure 4.9 UE/eNB Random Access Procedure State Diagram (cont.)

4.4.5 Attach and Default Bearer Setup

Following the random access procedure described in the previous section, The UE is able to send uplink data to eNodeB after reserving the needed resources and establishing the connection with eNodeB. In this subsection the attach process and default bearer establishment will be described in detail. After sending the RRC connection request message, eNodeB will be able to identify the UE by Cell Radio Network Temporary Identifier (C-RNTI) that was assigned to it during the random access procedure, then eNodeB responds with RRC connection setup message over DL-SCH in which the Signalling Radio Bearer (SRB) is created in Acknowledged Mode (AM), it also defines the power head room and uplink power control along with configuration parameters for uplink Radio Link Control (RLC) and UL-SCH.

Now the UE is ready for data reception and transmission as all the configuration parameters are set and initial resources in both uplink and downlink channels are reserved. The signalling sequence and context switching diagrams for attach and bearer establishment signalling is shown in figures 4.10 and 4.11. At can be listed as follows:

1) The UE uses the reserved UL-SCH resources to send RRC connection setup complete message, with attach request embedded as NAS payload, to eNodeB. The request contains the Globally Unique Temporary Identifier (GUTI) and the old Globally Unique MME Identifier (GUMMEI). eNodeB will be able to identify the MME using GUMMEI.

2) The attach message is forwarded by the eNodeB to the MME via S1AP interface. The message will also include the PDN connectivity request message, the Tracking Area Identifier (TAI), and the E-UTRAN Cell Global Identifier (ECGI).

3) The new MME will send Identification Request to the old MME (which was resolved from the GUMMEI) and use the GUTI of UE to get the International Mobile Subscriber Identity (IMSI) from old MME. The old MME responds with Identification response message embedded with IMSI and other information. Then the new MME send the ciphered Options request to the UE and get the response back in case that the UE has set the Ciphered Option Transfer flag.

4) The new MME sends Update Location Request message to the HSS to update it with the user current location identified uniquely by its IMSI and PLMN Identity. The HSS sends Cancel Location to the old MME which acknowledges the request and removes bearer context. Finally, HSS sends Update Location Request Answer to the new MME loaded with subscription data details retrieved from HSS such as EPS subscribed QoS Profile, Aggregate Maximum Bit rate (MBR) for both the downlink and uplink channels, APN, PDN GW address, QCI, and Charging details. The new MME validates the info and creates new context for the UE.

5) To start bearer establishment process, MME sends GPRS Tunnelling Protocol (GTP) Create Session Request message to the Serving Gateway (SGW). The request include the Quality Class Identifier (QCI) information, PGW IP address, PDN IP address, APN, IP Address assigned to UE, downlink and uplink MBR. The APN assigned by UE is used for default bearer activation (there could be multiple bearers for single UE).

6) The SGW creates a new entry in the EPS bearer table and makes a mapping between APN and related PDN GW address. Then SGW sends a request to create a default bearer to the PDN GW. The PDN in turn creates a new entry in its EPS bearer context table then it replies back to the SGW with Create Default Bearer Request Message. Now the PDN GW is able to route the user plane Protocol Data Units (PDUs) between the SGW and the packet data network.

7) The SGW receives and buffers the first downlink data block, it then sends a create default bearer request message to the MME which in turn sends to eNodeB a message that contains three embedded messages (S1AP initial context setup request, NAS attach accept, and activate default bearer request).

8) eNodeB will then extract and process each of the three messages consequently. It will first process the initial context setup request message that include MBR information for the UE, and QCI information. Both the uplink and downlink MBR along with QCI will be used by eNodeB to setup and allocate the needed resources for UE. Then it will extract and process attach accept message in which it signals the completion of the attach process. Then eNodeB will process the activate bearer request message that contains Radio Access Bearer (RAB) quality of service information, the Access Point Name (APN), and PDN address.

9) RRC connection reconfiguration message is then sent to UE to activate the default bearer, the UE activate the bearer then replies back to eNodeB with RRC connection reconfiguration complete message. Then eNodeB sends initial context setup response message, that includes eNodeB address to be considered for the downlink traffic, to the MME. UE sends attach complete message to eNodeB then it is forwarded to MME. Now the user can send data over uplink channels.

10) The MME sends then update bearer request message to SGW and gets the update bearer response message back from SGW. Now the SGW is able to forward the buffered downlink traffic down to the UE and forward the following downlink data to the UE without the need to buffering it.

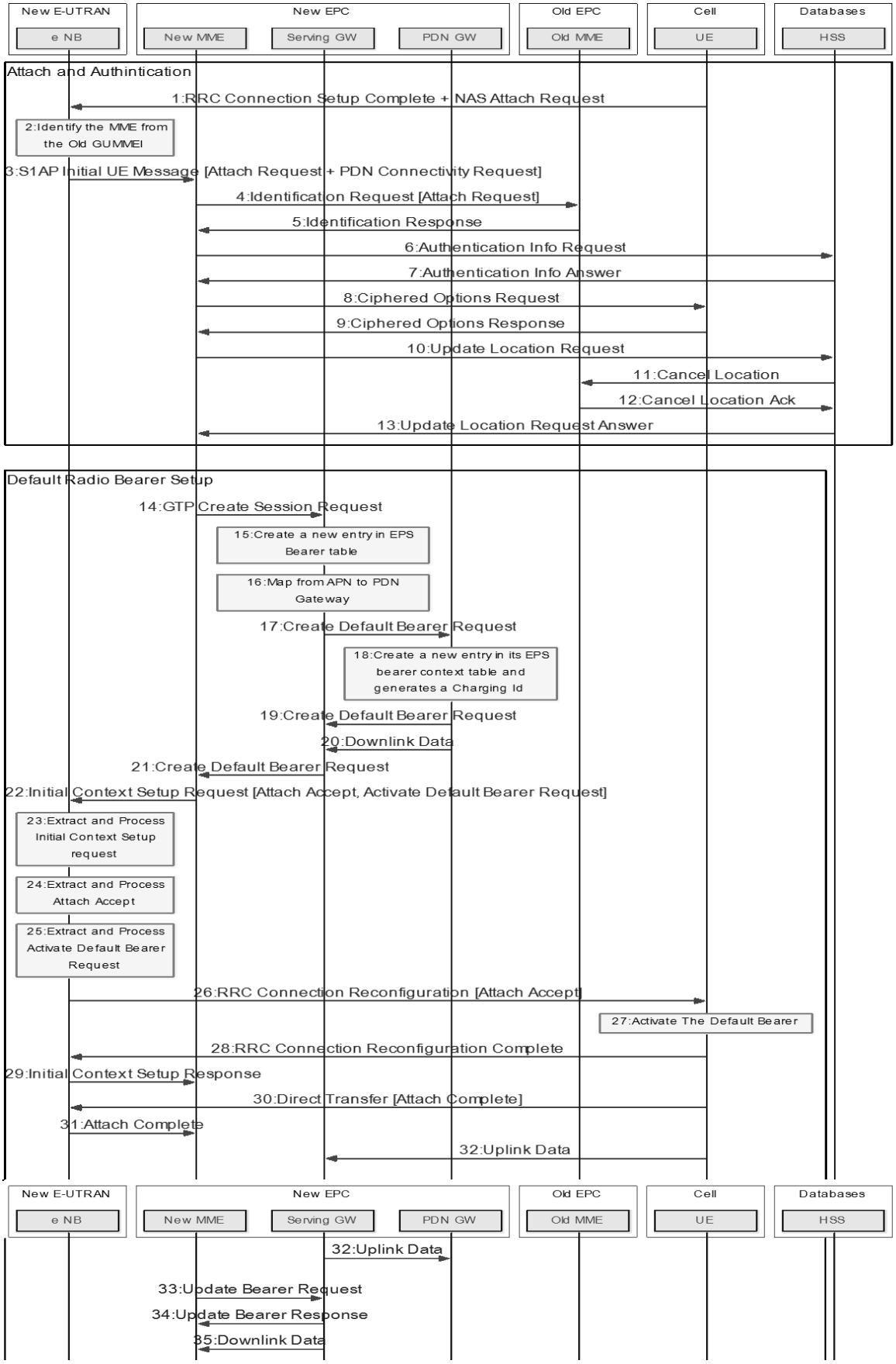


Figure 4.10 Attach and Default Bearer Setup Sequence Diagram

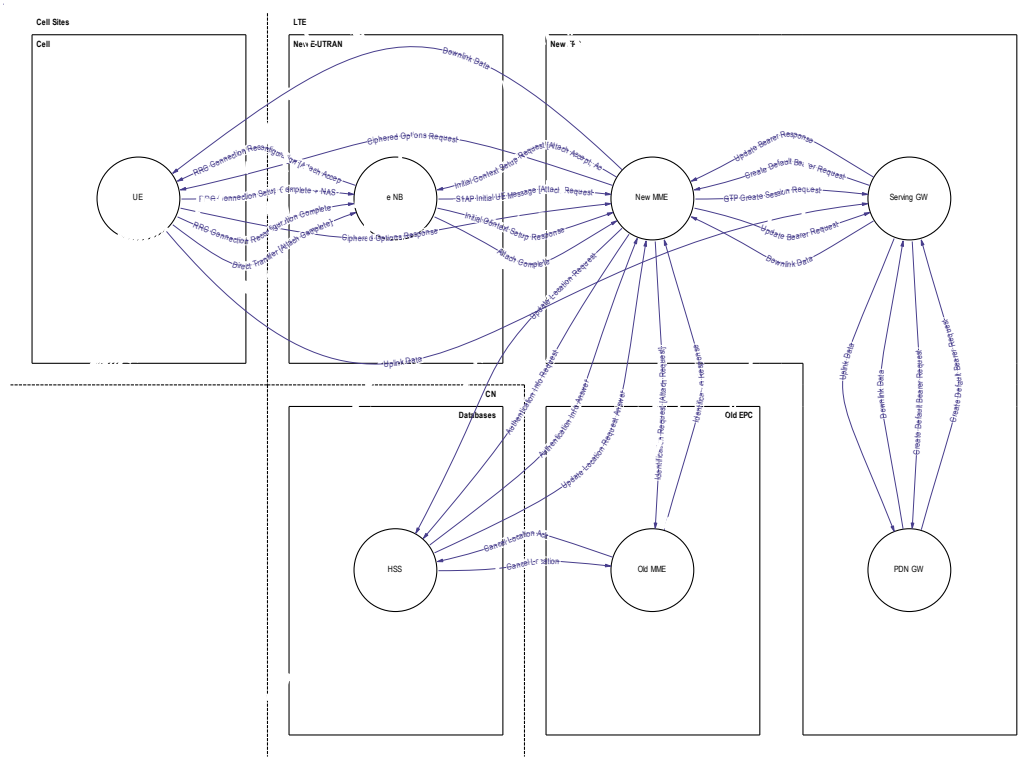


Figure 4.11 Attach and Default Bearer Setup State Diagram

4.4.6 VoLTE Signalling

In the previous section, it was shown that after finishing the attach and bearer setup process, the UE is able to send data over uplink control and data channels and receives downlink data over downlink control and data channels from eNodeB, it was shown that logical connections are established between MME and SGW and PDN GW with the needed mapping and routing tables that are able to monitor QoS requirements. VoLTE is the standard considered for voice call setup in LTE networks, it uses IMS signalling to setup the call with SIP servers and it has the advantage of enforcing certain QoS measure compared with the traditional best effort VoIP protocol. In this section the VoLTE signalling will be demonstrated to show the interaction with other entities in LTE U-TRAN, EPC, and IMS core Network.

3GPP has defined the logical architecture for VoLTE. LTE, EPC, and IMS core. The logical architecture is shown in figure 4.12

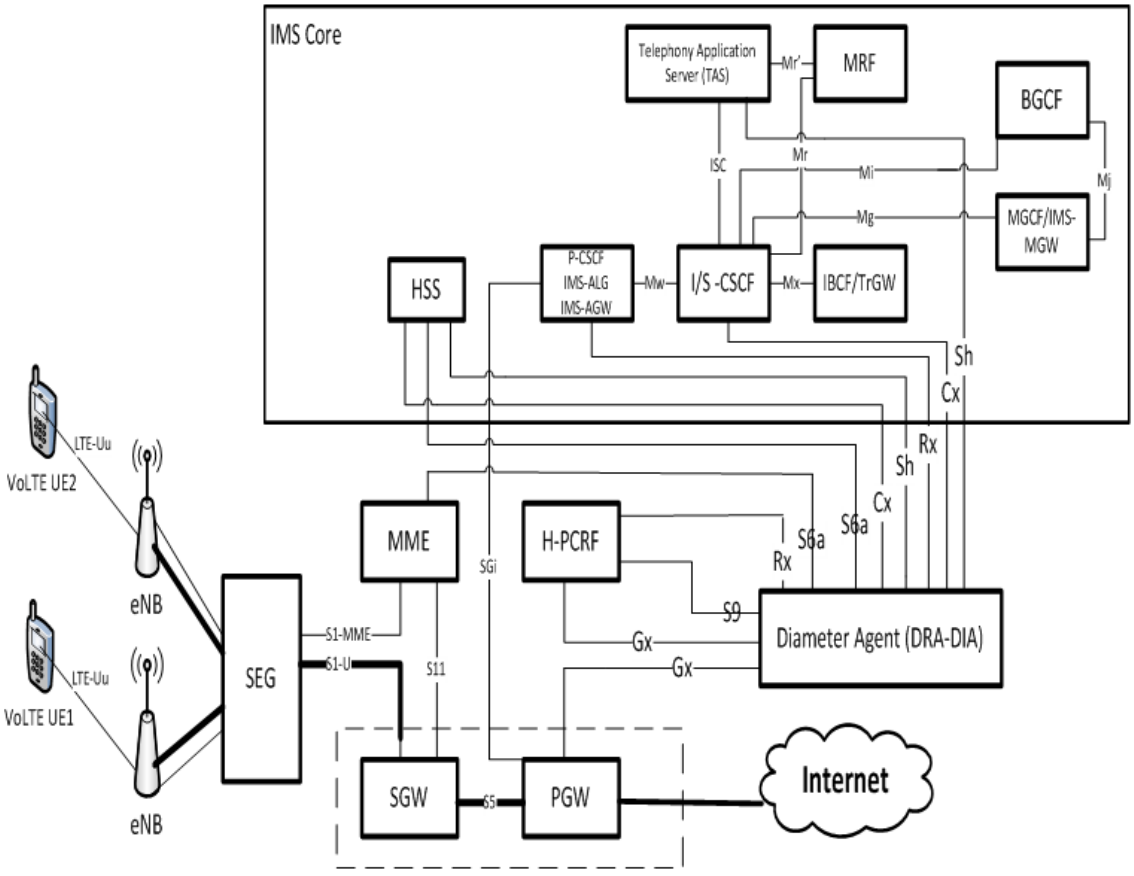


Figure 4.12 Volte Architecture

In order to show the LTE E-UTRAN, LTE EPC and IMS entities interactions. The VoLTE call (without roaming) message sequence for UE attach and IMS registration process sequence diagram and context switch are shown in figure 4.13. Similarly, the IMS call setup sequence diagram and context switch are shown in figure 4.14.

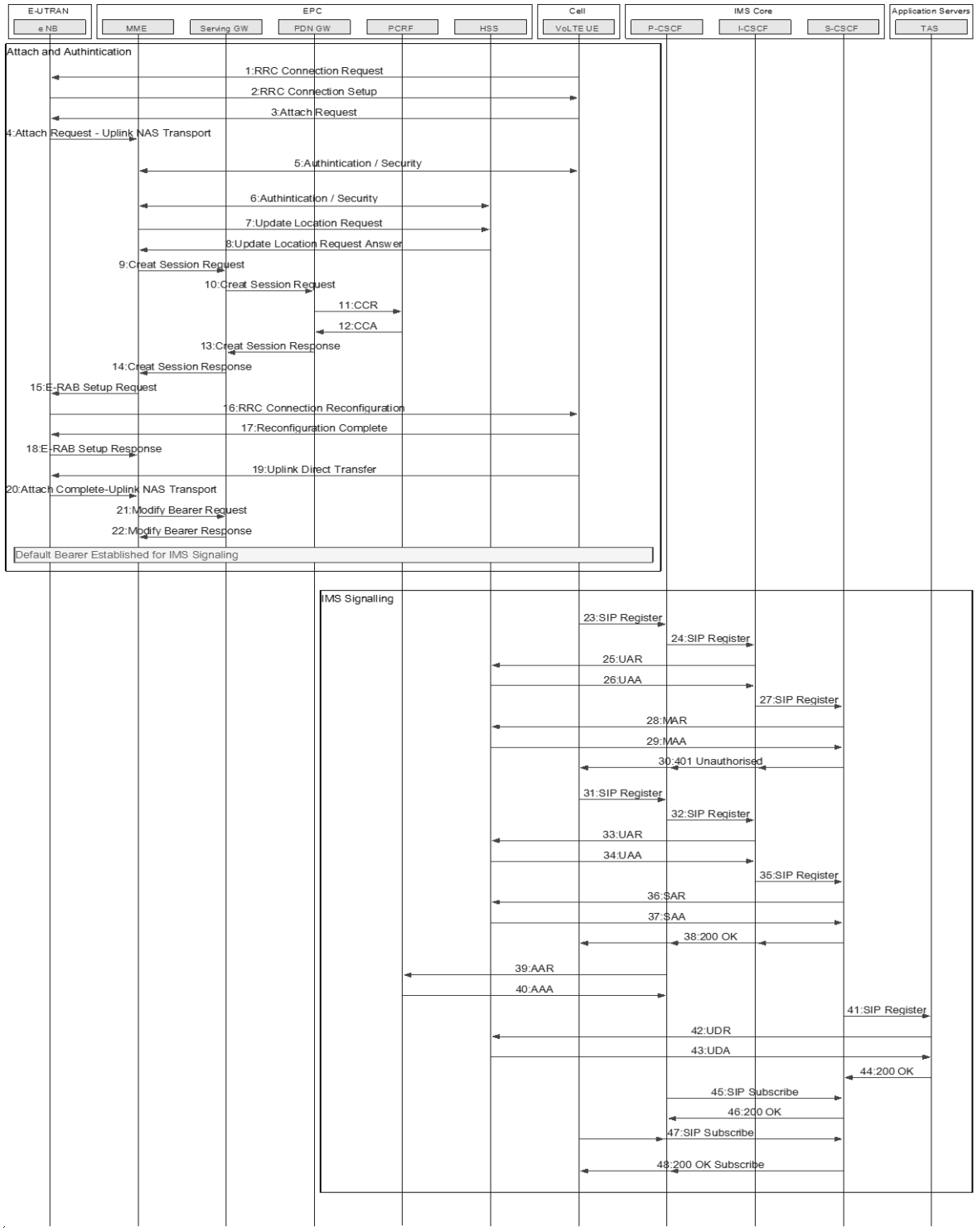


Figure 4.13 UE attach and IMS registration process Sequence Diagram

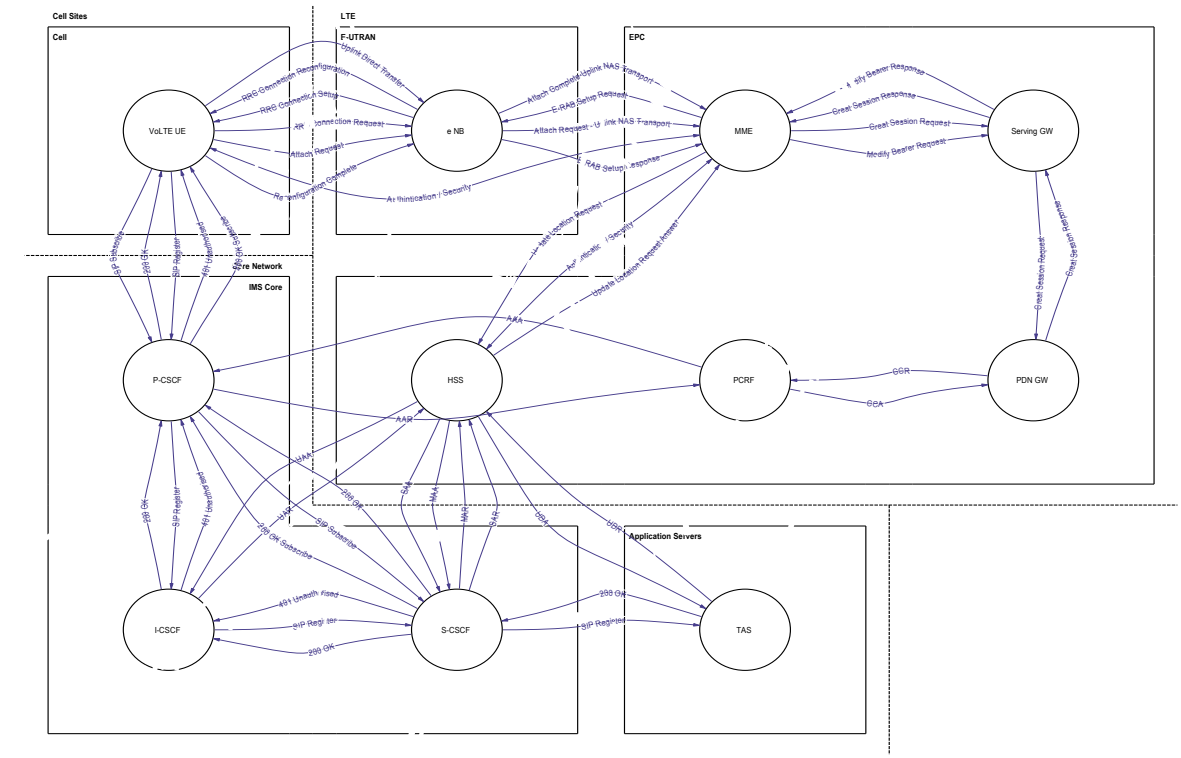


Figure 4.14 UE attach and IMS registration process State Diagram

4.4.7 IMS call setup

Following the attach and default bearer activation process described in the previous section, the IMS client operation in VoLTE UE is ready now to establish IMS call setup signalling, P-CSCF is discovered and S-CSCF is identified by UE during the previously described registration process. Figures 4.15, 4.16, 4.17, and 4.18 show the sequence diagrams for both the origination and terminating UEs signalling flow along with the involved intermediate entities. For simplicity, both UEs are assumed to be home users (both are not roaming) and the EPC and core IMS entities belong to the same Public Mobile Network (PMN) in which the same entities are accessed by both UEs. For simpler handshaking presentation, the sequence diagrams for both the origination and termination sides are shown separately, similarly the context diagrams for both entities are displayed separately. However the timely interactions will be described below for both entities sequentially according to the occurrence order.

The IMS call procedure can be summarised as follows:

1) VoLTE UE1 initiates SIP INVITE message, with SDP offer of IMS capabilities and preferred codec, and sends it to P-CSCF. The packet travels through the previously established default bearer through eNodeB, SGW (Serving Gateway), PGW (Packet Data Network Gateway), and finally to P-CSCF. Then P-CSCF adds its own header and forwards it to the S-CSCF. S-CSCF check the subscriber profile it has got during the registration process, then it checks the P-Preferred-Service header in the invite message (MMTel ICSI) and check the caller profile if the needed service is among the subscribed services. S-CSCF then forwards the invite request to TAS to invoke VoLTE supplementary services, TAS then invoke the supplementary services logic and forwards invite request to S-CSCF. S-CSCF check the callee address and decides that the callee is within its home network (either via Telephone Number Mapping (ENUM)

/DNS lookup or via internal configuration since both users have previously registered and been recognised), S-CSCF check the callee subscriber user profile now and forwards the request to P-CSCF and then to the callee VoLTE UE2.

2) VoLTE UE2 (Callee) replies with SDP answer in a SIP 183 Progress message. The SDP answer should indicate the preferred codec approval as requested by the caller but the QoS requirements are not yet met at the terminating site as there is no dedicated bearer established yet. S-CSCF gets the SIP 183 progress message and forwards it to P-CSCF. P-CSCF analyse the SDP answer with SIP 183 message and send Authorise/Authenticate request to Policy and Charging Rules Functions (PCRF) with all related service related information to check the charging profile, the service related information include IP addresses, port addresses, and media type. The PCRF authorise the request and map it with the list of user subscription list of allowed services and QoS information. Then PCRF identifies the affected default bearer and then initiates Re-Auth-Request (RAR) to PDN GW to create a dedicated bearer for voice data along with QoS profile (with QCI=1, ARP)

3) Following RAR reception by PDN GW, it forwards create bearer request message over the pre-established default bearer channel dedicated for control data signalling, to Serving GW which in turn forwards the request to MME. MME sends activate dedicated bearer request to eNodeB which then forwards RRC connection reconfiguration message to the UE, the UE stores the dedicated bearer identity and links it with the default bearer replies activate dedicated bearer accept message to eNodeB and following the same path back the activate bearer response message sent to MME and create bearer response is sent to Serving GW and then forwarded to PDN GW. Finally, PDN GW sends Re-Authentication-Answer (RAA) to PCRF, then PCRF sends Authorise-Authenticate-Answer (AAA) message back to P-CSCF.

4) Now the dedicated bearer for voice data is established. P-CSCF then forwards SIP 183 Progress message to VoLTE UE (Caller) which then send to the Callee a Provisional Response Acknowledgment (PRACK) with an associated 200 OK (PRACK) to indicate the successful reception of 183 Progress message. It then confirms the reservation of local resources and the selected codec and to show that all preconditions has been met and the media stream is active due to the establishment to dedicated bearer. The PRACK is sent first to P-CSCF and S-CSCF then to the callee side. The caller then gets 200 OK (UPDATE) response from the Callee to indicate that a codec has been selected and the resources are reserved and the media stream is activated similar to the way the caller has done before.

5) After meeting the preconditions at both the Caller and Callee sides, the callee will be alerted upon reception of 200 OK (UPDATE) message and it then will send 180 Ringing message to the caller through S-CSCF and P-CSCF. The Caller UE will generate a local ring tone to be heard by the subscriber. Once the call is picked up by the Callee, it sends 200 OK message to the caller through S-CSCF and P-CSCF. P-CSCF will invoke PCRF by sending AAA message to enable downlink and uplink of the dedicated bearer. PCRF then sends RAR message to PDN GW to enable the media flow through it. RAA sent back to PCRF and AAA is sent back to P-CSCF to acknowledge the channel enable process. P-CSCF then forwards 200 OK (INVITE) message to the caller which reply with SIP ACK message to indicate that the call has been established and then the voice RTP traffic between end users will be sent over the dedicated bearer, whereas the control signalling will be exchanged using default bearer channel.

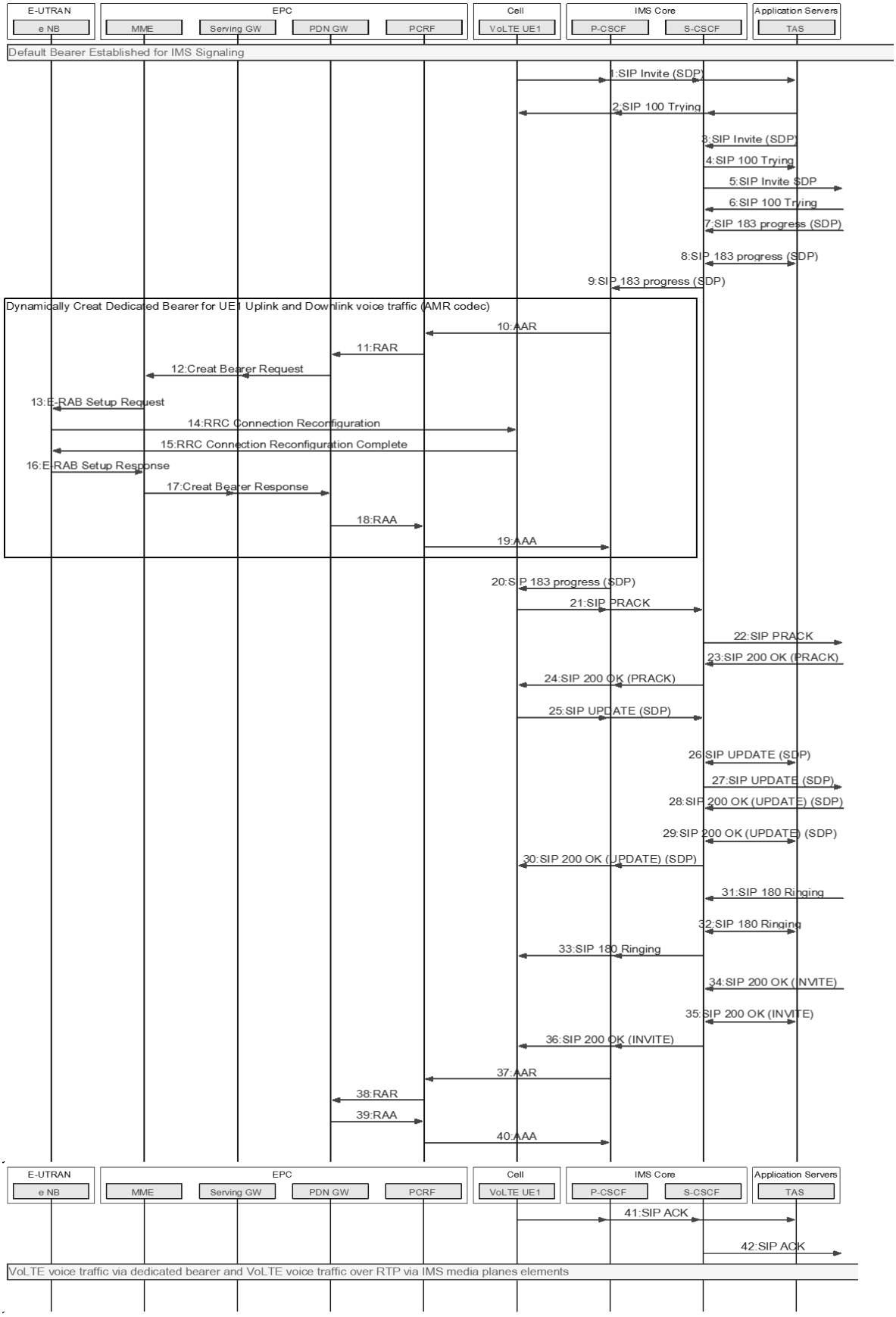


Figure 4.15 IMS call setup for origination UE Sequence Diagram

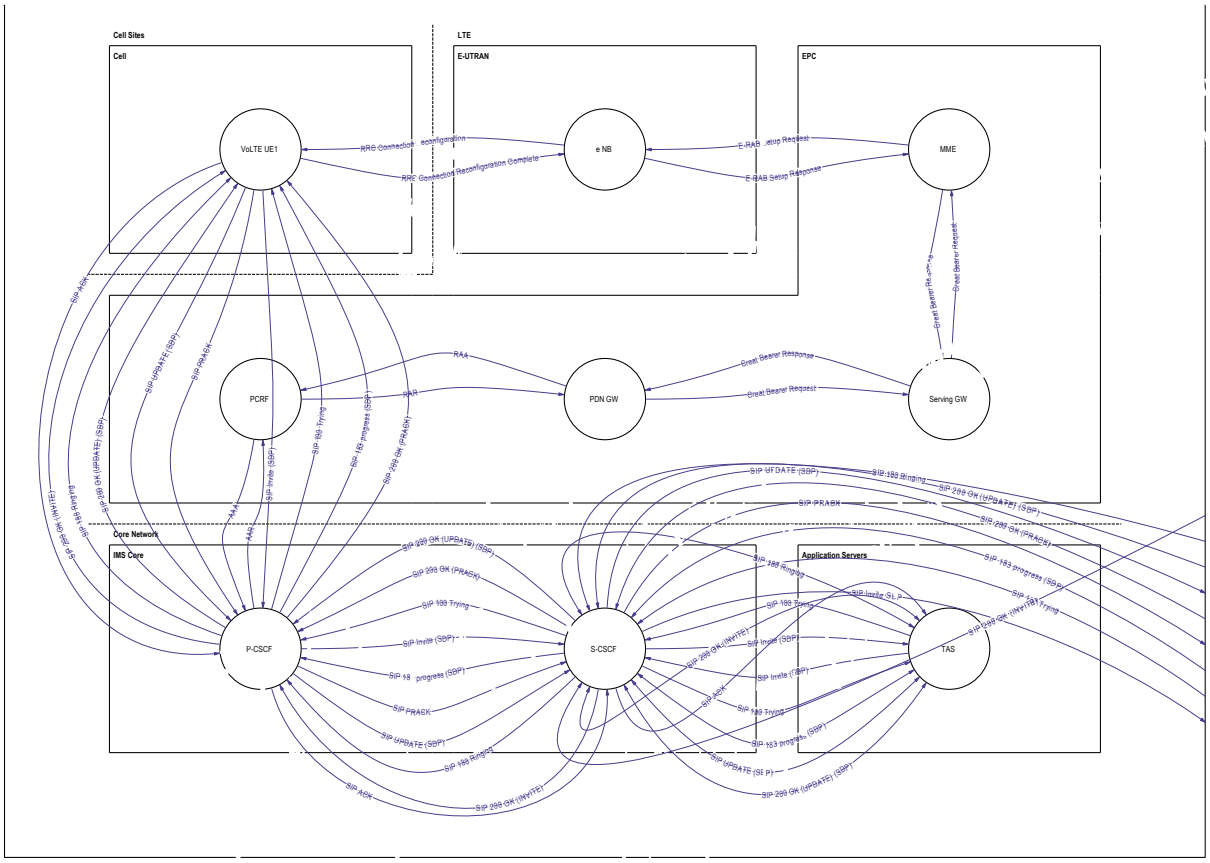


Figure 4.16 IMS call setup for originating UEs State Diagram

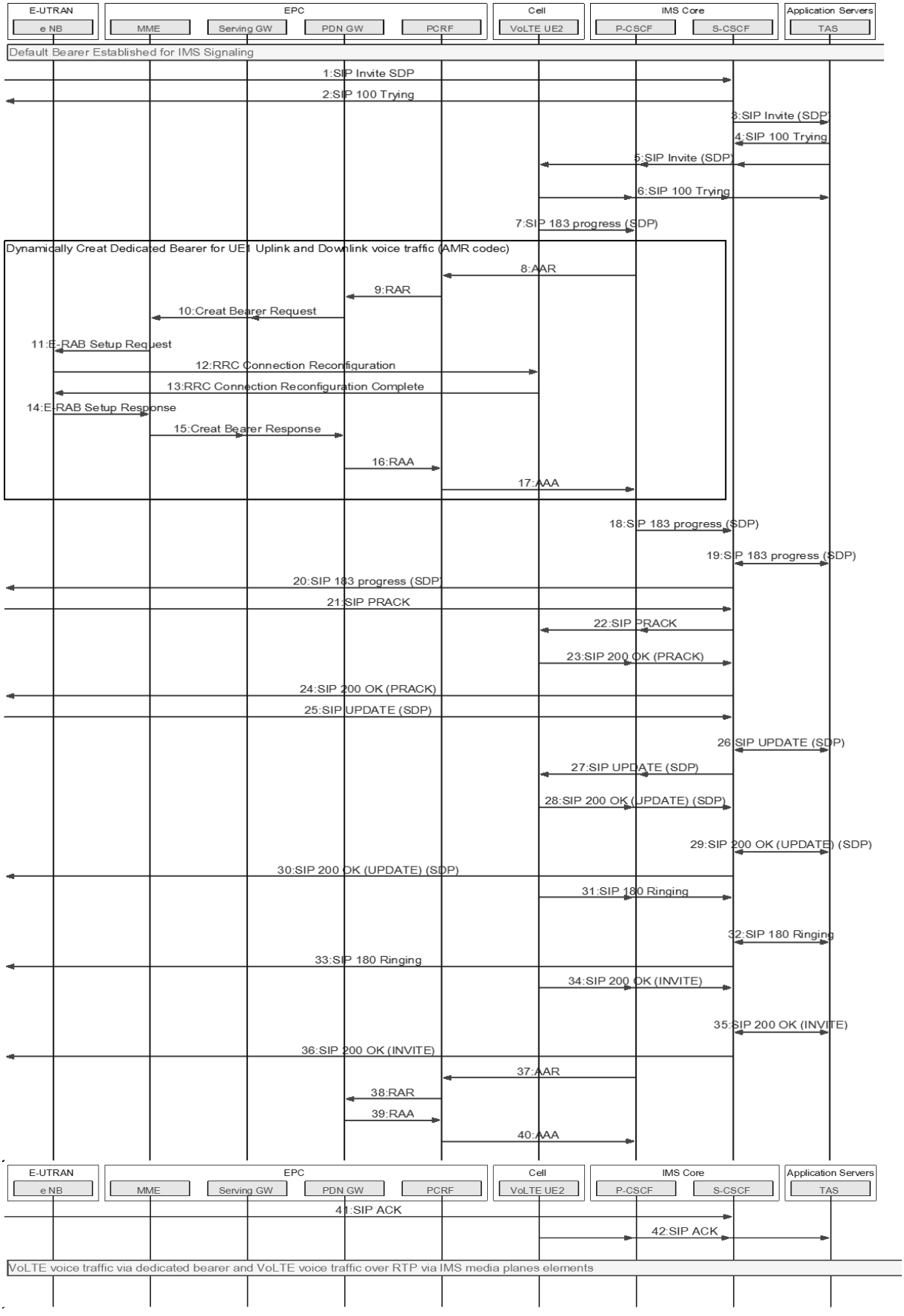


Figure 4.17 IMS call setup for terminating UEs Sequence Diagram

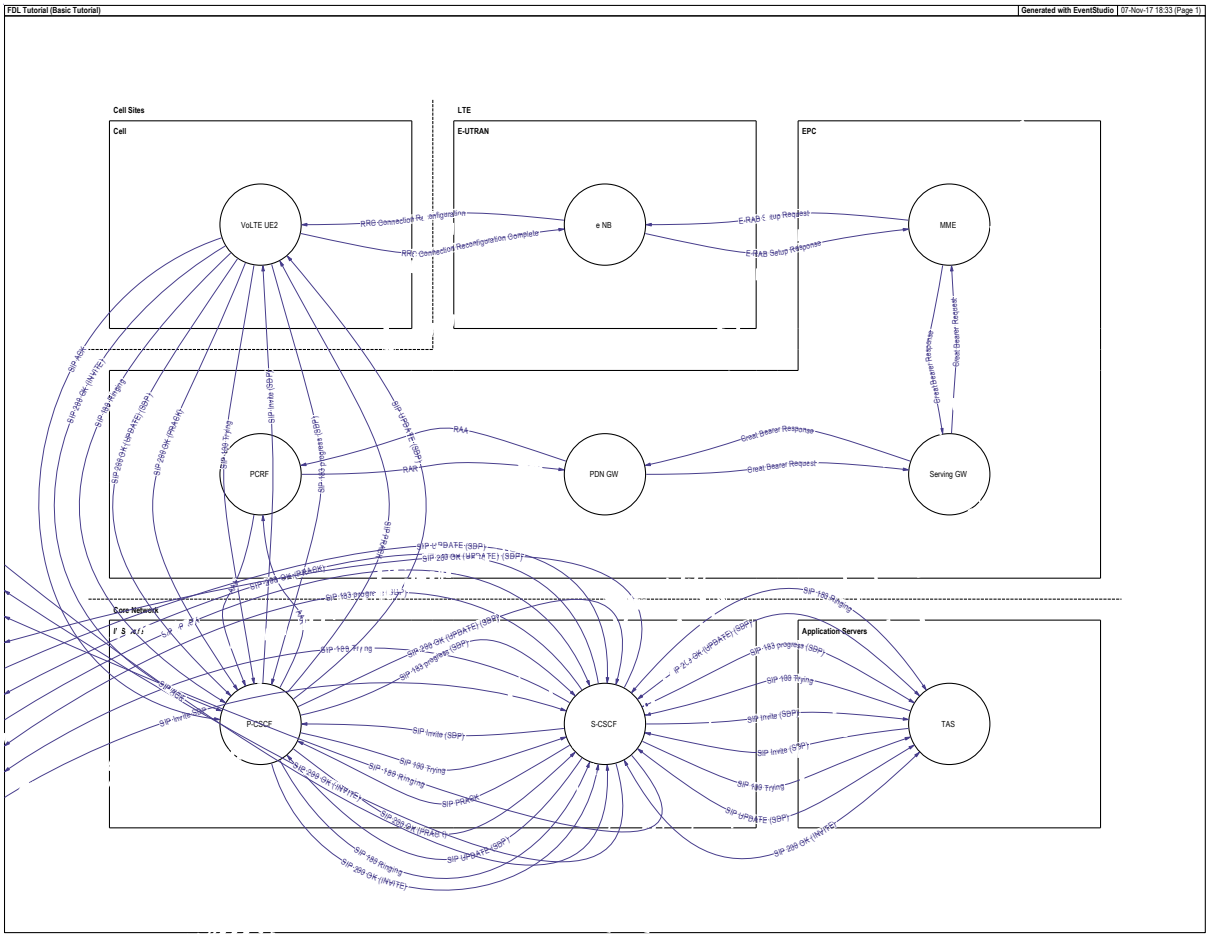


Figure 4.18 IMS call setup for terminating UEs State Diagram

4.5 BOTTLENECK ANALYSIS

4.5.1 Methodology

In this section, the bottleneck analysis of the communication system will be studied to evaluate the overall system availability and scalability. Based on the overall system framework interconnections shown previously and following the signalling flow pattern of the sequence diagram shown in the previous sections, it is possible to generate statistical estimations of the entities load by counting the number of hits for each server.

The entities can be classified based on the organisational hierarchal structure into different categories; objects, modules, and components. Figure 4.19 shows the classification of entities/subsystems based on the aforementioned three categories.

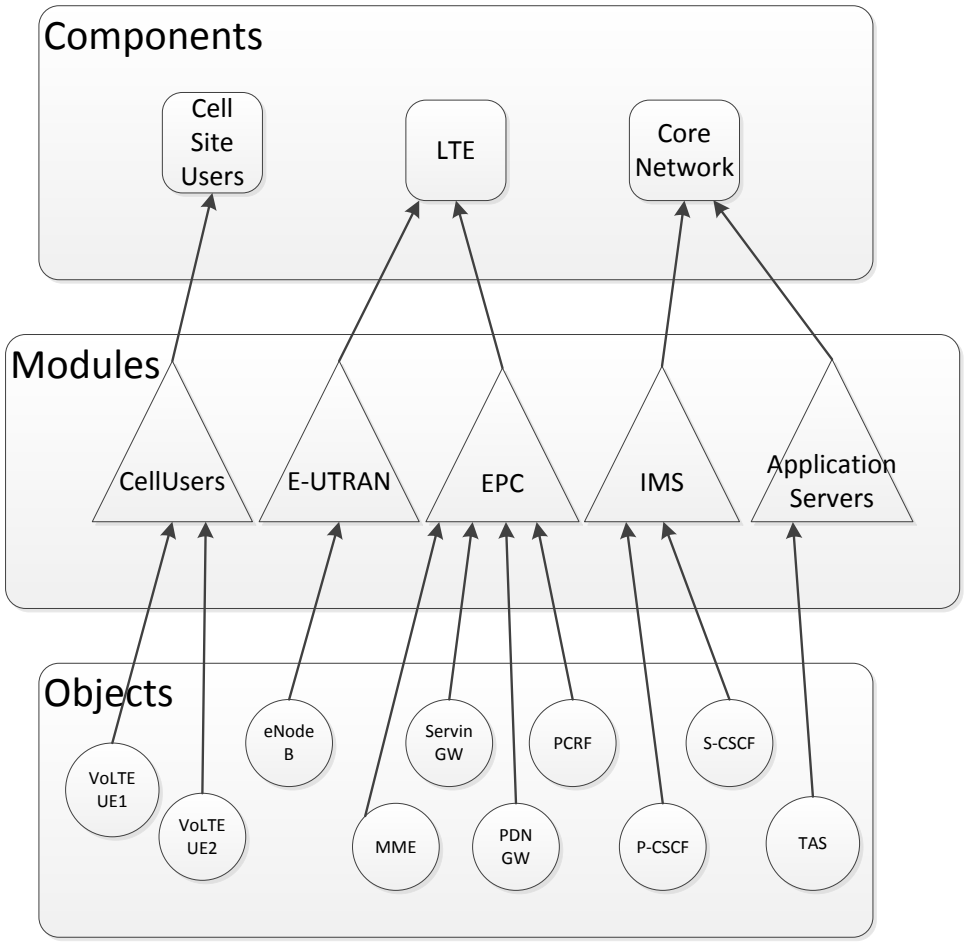


Figure 4.19 Entities Organisational Hierarchal Structure

4.5.2 Analysis

Following the organisational structure in figure 4.19. It would be easier to calculate the number of hits at the object level, module level, and component level to get better estimate of the bottleneck and load distributions. Using simple .Net Programming code, detailed statistics was carried on accordingly.

Invitation Process Signalling Analysis

Table 4.1, table 4.2, table 4.3, table 4.4, table 4.5 and table 4.6 summarise the statistics collected for the call invitation signalling at both the caller (VoLTE UE1) and callee (VoLTE UE2) sides. It is important to mention that both the Callee and the Caller are using the same access network and served by the same core network elements, so the same IMS entities are serving both VoLTE UE1 and VoLTE UE2, it is also assumed that both users are at their home network. This realistic assumptions avoid the signalling complexities introduced due to user roaming between different networks and it also avoids introducing the overhead of forwarding the signal to different IMS domain.

Table 4.1 Caller side interactions

*	Cell Sites	LTE	Core Network
Cell Sites		1	7
LTE	1	10	2
Core Network	3	2	15

Table 4.2 Callee side interactions

	Cell Sites	LTE	Core Network
Cell Sites		1	4
LTE	1	10	2
Core Network	6	2	15

Table 4.3 Caller side inter module interactions

*	Cell	E-UTRAN	EPC	IMS Core	Application Servers
Cell		1		4	
E-UTRAN	1		1		
EPC		1	8	2	
IMS Core	6		2	10	3
Application Servers				2	

Table 4.4 Callee side inter module interactions

*	Cell	E-UTRAN	EPC	IMS Core	Application Servers
Cell		1		7	
E-UTRAN	1		1		
EPC		1	8	2	
IMS Core	3		2	10	3
Application Servers				2	

Table 4.5 Caller side Inter object interactions

*	eNB	MME	Serving GW	PDN GW	PCRF	VoLTE UE1	P-CSCF	S-CSCF	TAS
eNB		1				1			
MME	1		1						
Serving GW		1		1					
PDN GW			1		2				
PCRF				2			2		
VoLTE UE1	1						4		
P-CSCF					2	6		4	
S-CSCF							6		3
TAS								2	

Table 4.6 Callee side inter object interactions

*	eNB	MME	Serving GW	PDN GW	PCRF	VoLTE UE2	P-CSCF	S-CSCF	TAS
eNB		1				1			
MME	1		1						
Serving GW		1		1					
PDN GW			1		2				
PCRF				2			2		
VoLTE UE2	1						7		
P-CSCF					2	3		7	
S-CSCF							3		3
TAS								2	

Registration process Signalling Statistics

Similar to the call setup analysis presented in the previous section, the registration signalling statistics are collected and listed in table 4.7, table 4.8, and table 4.9. In which one of the users only has been considered. It is important to note that all users are registering their details in the network separately and independently and at different time intervals. So there is no need to show the signalling statistics for both users as the other user will have exactly the same results. In contrary to the call establishment results presented before, the I-CSCF is now involved in the registration signalling.

Table 4.7 Registration process inter module message interactions

*	Cell Sites	LTE	Core Network
Cell Sites		4	3
LTE	2	14	6
Core Network	3	6	14

Table 4.8 Registration Process inter component message interactions

*	Cell	E-UTRAN	EPC	IMS Core	Application Servers
Cell		4		3	
E-UTRAN	2		3		
EPC		1	10	5	1
IMS Core	3		5	12	1
Application Servers			1	1	

Table 4.9 Registration process inter object message interactions

*	eNB	MME	Serving GW	PDN GW	PCRF	HSS	VoLTE UE	P-CSCF	I-CSCF	S-CSCF	TAS
eNB		3					2				
MME	1		2			1					
Serving GW		2		1							
PDN GW			1		1						
PCRF				1				1			
HSS		1							2	2	1
VoLTE UE	4							3			
P-CSCF					1		3		2	2	
I-CSCF						2		2		2	
S-CSCF						2		2	2		1
TAS						1				1	

The tables shows the load based on the traffic direction (from source to destination) in which rows represent sources and columns represent destinations. From tables 4.1, 4.2, and 4.7, it is shown that during both registration and invite signalling the core network components experience the highest number of traffic hits which reflect the highest load of SIP traffic in the communication chain. Tables 4.3,4.4,4.8 show that IMS is the hub of most of the core traffic, this may look obvious as it is the only entity that process and forwards the SIP traffic. However, extending it further to the more detailed statics of IMS entities statistics shown in tables 4.5, 4.6, and 4.9, it becomes clear that P-CSCF and S-CSCF servers are the bottleneck for this scenario and the aforementioned assumptions.

When it comes to providing a reliable and more scalable solutions that are supposed to be considered as a backbone system for mission critical users, System availability and reliability under high load pressures is mainly the core requirement need to be sustained at all times. The next sections will demonstrate the methodology followed to study the core IMS components performance and enhancement approaches.

4.6 PERFORMANCE EVALUATION METHODOLOGY AND TOOLS

The available tools and the followed methodology will be described in this section. The methodology followed to perform the experiments relies on the available tools, evaluation of results need to be considered according to a reference design. The research methodology since the early stages of research progress can be summarised as shown in figure 4.20. Different evaluation tools have been used according to the implemented task and expected outputs for certain stage. A brief description for each stage will follow in the following subsections leaving the detailed implementation and results for the next chapter.

Following the organisational structure in figure 4.19. It would be easier to calculate the number of hits at the object level, module level, and component level to get better estimate of the bottleneck and load distributions. Using simple .Net Programming code, detailed statistics was carried on accordingly.

Invitation Process Signalling Analysis

Table 4.1, table 4.2, table 4.3, table 4.4, table 4.5 and table 4.6 summarise the statistics collected for the call invitation signalling at both the caller (VoLTE UE1) and callee (VoLTE UE2) sides. It is important to mention that both the Callee and the Caller are using the same access network and served by the same core network elements, so the same IMS entities are serving both VoLTE UE1 and VoLTE UE2, it is also assumed that both users are at their home network. This realistic assumptions avoid the signalling complexities introduced due to user roaming between different networks and it also avoids introducing the overhead of forwarding the signal to different IMS domain.

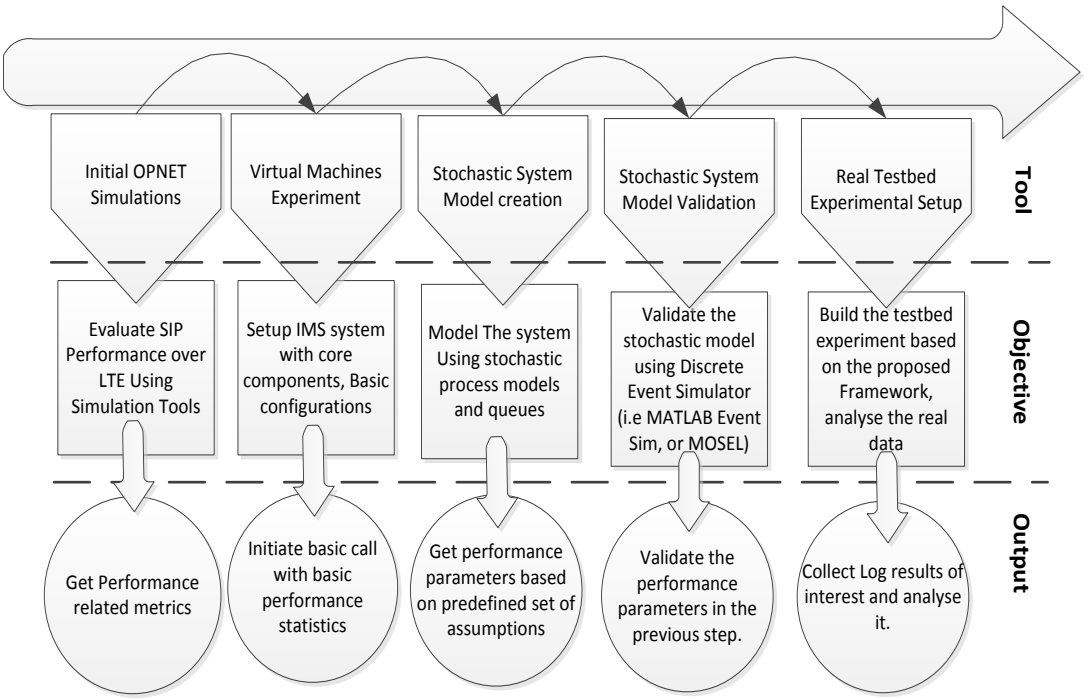


Figure 4.20 Tools selection Methodology

4.6.1 Initial OPNET Simulations

OPNET as a Discrete Event Simulator (DES) is widely used by the research community as a network testing, configuration and engineering tool. During the early stages of the research, the availability of the tool was utilised to carry out initial performance evaluation of simple SIP server performance over 4G LTE access technology. It was part of defining the gap in the research part during the research progress plan. Defining the most important performance metrics that are related to general SIP signalling is the one of the main goals of this stage, a more detailed analysis of the results will follow in the next chapter.

4.6.2 IMS Experiment over Virtual Machines

Following the simple SIP server tests, IMS testbed experiment over virtual machines was implemented in which the basic functionality of the basic IMS core components was tested and evaluated. IMS, as a set of interconnected SIP servers, has multiple interfaces and databases that interact with each other using mainly SIP signalling. The basic functions performance was monitored and the configuration parameters related to SIP performance according to the

metrics defined in the earlier stage was set. More details of the setup will be described in next chapter.

4.6.3 System Model Creation

While investigating the literature related to SIP performance, it was found that many studies in the literature has implemented different system model according to different queuing theory models to come up with a near real approximation of real systems. Therefore, along with the experimental testbeds, a system model that predict the performance of SIP metrics is implemented. The proposed model serves as an integral part of the contribution and as a reference benchmark of SIP performance based on predefined metric measures.

4.6.4 System Model Validation

The proposed model in the previous stage was tested and validated using one of the stochastic models simulators to validate the assumptions set for the model and to visualise and measure the system response and behaviour against different loads. The load effect over the model and its influence over the output will be analysed to get better understanding of the theoretical scalability of the proposed solution.

4.6.5 Real-Time Testbed Experimental Setup

At the final stage, the testbed according to the proposed framework design is built and set. Results is collected and analysed according to the predefined measures of interest and following the guidelines of benchmarking recommendations found in the standard reports and literatures.

4.7 SUMMARY

This chapter implemented a systematic methodology helped in identifying the bottleneck point for different signalling types. According to the statistics collected for registration and invitation signalling, both the access and core networks experience heavy signalling overhead for serving one user. To have a scalable solution as suggested in the previous chapter, an analysis for possible bottleneck joints in the network. Serving large amount of user may fail part of the network entities. To achieve this goal, a set of tools was used in systematic way to design a framework and set the basis of the evaluation process that will be presented in the next two chapters.

Following the organisational structure presented in this chapter the number of hits at the object level, module level, and component level were calculated to get better estimate of the bottleneck and load distributions. Using simple .Net Programming code, detailed statistics was carried on accordingly. The collected statistics for the call invitation signalling at both the caller (VoLTE UE1) and callee (VoLTE UE2) sides showed that there is nodes that experience more traffic than other nodes.

It is shown that during both registration and invite signalling the core network components experience the highest number of traffic hits which reflect the highest load of SIP traffic in the communication chain. Similarly, it was proven that IMS is the hub of most of the core traffic, this may look obvious as it is the only entity that process and forwards the SIP traffic. However,

extending it further to the more detailed statics of IMS entities statistics, it becomes clear that P-CSCF and S-CSCF servers are the bottleneck for IMS.

When it comes to providing a reliable and more scalable solutions that are supposed to be considered as a backbone system for mission critical users, System availability and reliability under high load pressures is mainly the core requirement need to be sustained at all times.

Chapter 5: ENHANCED IMS FOR MCS FRAMEWORK DESIGN

5.1 INTRODUCTION

Following the signalling analysis presented in the previous chapter, this chapter will propose a framework designed as a new enhanced IMS system to be deployed in the core network. The newly proposed solution should overcome the scalability and system responsive issues highlighted in the previous chapters. The chapter will start with the design methodology, then it will present the design model and profile of the system, and finally it will end with the description of the detailed system intended functionality that will be described via flowcharts and algorithms.

5.2 FRAMEWORK DESIGN METHODOLOGY

The proposed framework design is based on the currently deployed IMS system performance in the market. According to the literature, there are challenges associated with the current design when it comes to providing a scalable, more reliable and responsive solution in the core network. The methodology followed during the framework design stage takes into account the 5G requirement of future network system that fits with future mission critical system's needs. Figure 5.1 shows the organizational diagram of the framework design methodology structure.

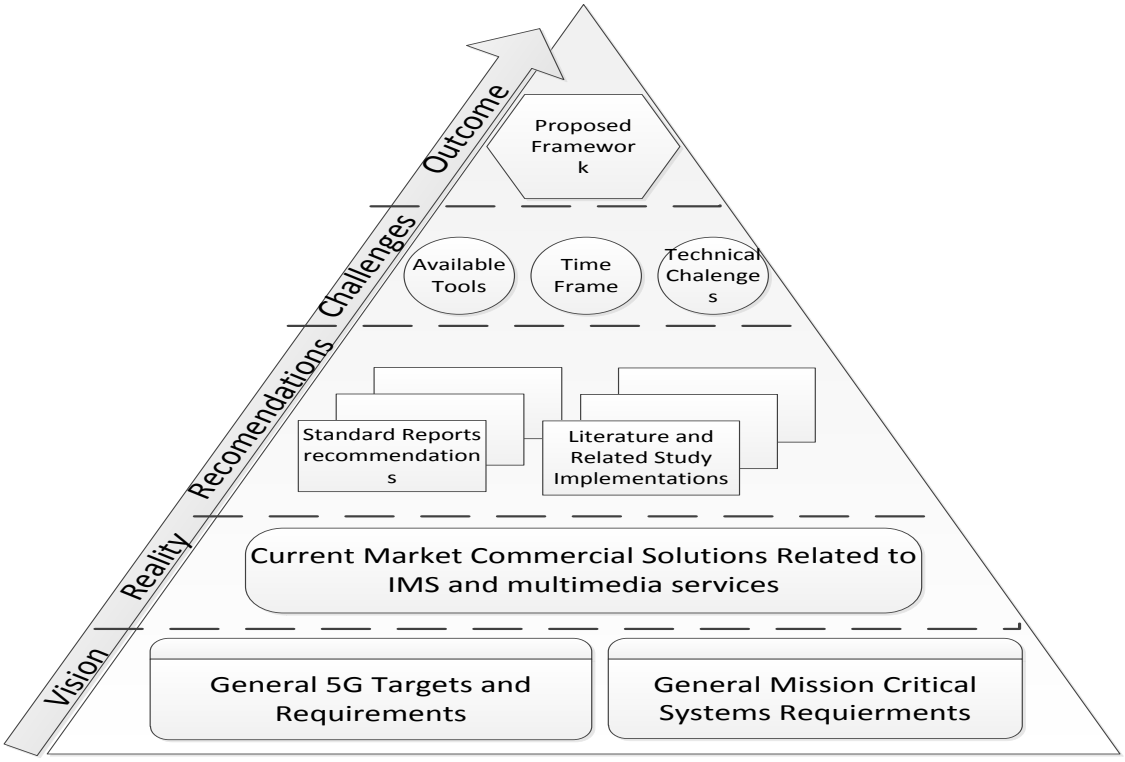


Figure 5.1 Organisational Diagram of the Framework Design Methodology

The structure represent the logical flow of the followed methodology to design the framework. The vision of the forthcoming enhancements multimedia services in the core network was introduced under the context of 5G Networks and future Mission Critical Systems recommendations. Service requirements investigated, for both 5G and MCS, has been used as a basis for any proposed solution. The future vision tend to be in general very optimistic and idealistic that in certain occasions collide with many constrains, a study of current solutions of multimedia services in the market and it performance drawbacks gives more realistic vision of current technologies and solutions capabilities. This non idealistic, closer to reality, approach pushed the research toward narrowing the gap of the field. Research efforts done by the research community, summarised in the literature review chapter, helped in identifying the general research trend and efforts paid toward filling the identified gap. Similarly, the technical and standard reports provided guidelines for research community to setup the research plans based on generalised methodological approaches. challenges has shaped the way the framework was designed by trying to exploit the current available resources and evaluation tools within the allocated time frame and according to the individually equipped or gained technical experience.

5.3 FRAMEWORK DESIGN MODEL

The Framework provides a description of the interconnections between different entities of the proposed system. It is up to the researcher to decide the suitable implementation and evaluation method in hardware, software, or a combination of both of them. To make it easier for the reader, the high level framework model is split into three parts as shown in figure 5.2.

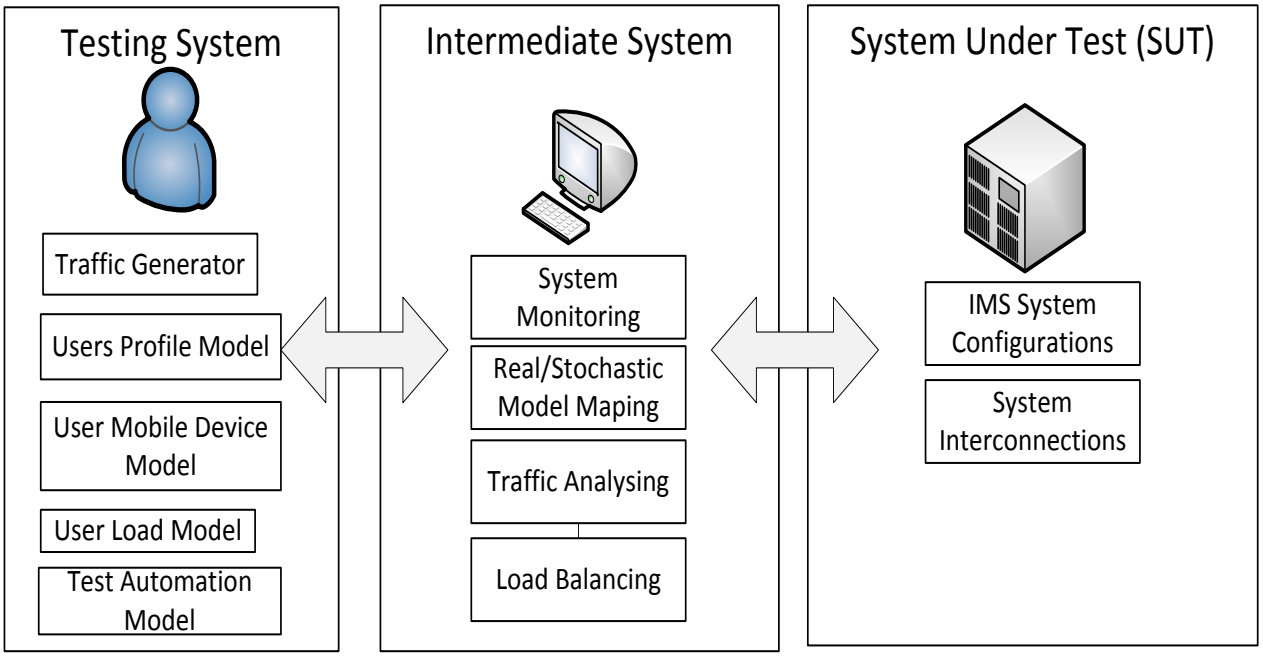


Figure 5.2 High level framework model

5.3.1 Testing System

The testing system models the load being generated to the System Under Test (SUT). The test traffic is supposed to replicate real world scenario traffic characteristics and parameters. On contrary to simple general purpose traffic or packet generators, the Testing system is designed to target the SUT by generating well shaped traffic that reflects different scenarios under different conditions. For this task, the testing system need to replicate the user's profile, number of users, user's mobility, user's mobile terminal characteristics, single user load model, and accumulative massive users load model.

In (ETSI 2010), ETSI has proposed a methodology and framework for testing and specification (MTS) of automated interoperability testing. The main objective of the report is to provide a generic framework and methodology for testing the standards regardless of the vendor, network, or service environment. Automation is introduced by ETSI as integral part of testing mechanism, there are different degrees of automation process; text execution, generation of test reports, computation of test verdicts, monitoring of relevant interfaces, etc., are all examples of automation level and degree which depends on the test execution context.

In another more specified document, ETSI has defined in (ETSI, 2010) the IMS related details for interoperability test automation recommendations. IMS, as a multimedia service control platform, relies on IP and SIP protocol to create isolation between the application and access layers via working as middle horizontal layer between them. Therefore, IMS does not standardise specific applications in the application layer nor demand certain underlying access technology standard interfaces to be integrated to its service. Due to the multi-standards multi-vendor generic functionality and architecture nature of IMS operation, the need for interoperable and more generic testing methodology to ensure meeting the test targets regardless of the implementation and detailed specifications of the equipment and standard being used.

An architecture of the framework and methodology for IMS core components was presented by ETSI in (ETSI 2010). An example interoperability test system configuration for IMS network to network interface was provided. The architecture deals with core IMS components as a black box and introduced multiple monitoring nodes for the interfaces of interest all linked to central monitoring coordination point that manages the carried test logic. In (ETSI 2013) the interoperability between IMS and EPC core entities is described.

5.3.2 Intermediate System

The intermediate system is the mail entity that sit in the middle between the testing unit and the SUT unit. Although logically positioned in between two system, it is designed to not interfere with the normal operation between the traffic source (Testing System) and the traffic sink (SUT). Normally, there should no added overhead or burden to the system to have a better estimate of the system performance, the intermediate system is therefore designed to monitor and process the sniffed traffic in both directions with minimal or no delay introduced.

The intermediate system functions can be summarised as follows:

1. Monitor and sniff the traffic generated from the testing system, the testing system represents an equivalent abstract of the underlying layers traffic.
2. Monitor and sniff the traffic generated from the SUT. Multiple implementation of SUT will be detailed in the next subsection.
3. Log the sniffed traffic in both directions and analyse it accordingly after applying certain predefined filtering criteria.
4. Hold a predefined system model information to be referred to occasionally during the analysis stage.
5. Compare the real traffic analysis data with the model-based reference data and analyse the performance accordingly.
6. Apply the SUT load balancing logic after analysing the related data parameters and leave the implementation of load balancing mechanism to the SUT subsystem which will be discussed in the next subsection.

According to the previously described functionality of the intermediate system, figure 5.3 shows the interconnections and block diagram of the different components embedded within the architecture of the proposed intermediate system.

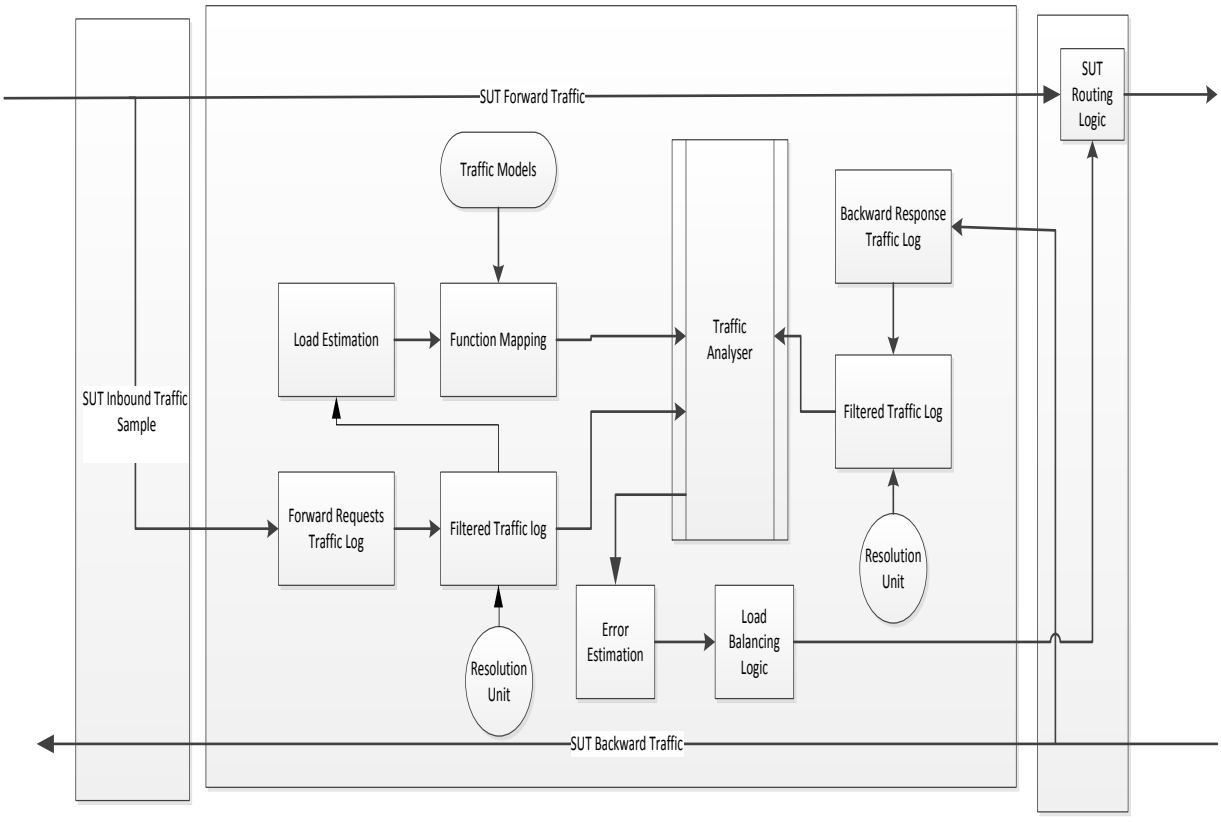


Figure 5.3 Intermediate System Interconnections

It is important to mention that all the block diagrams can be implemented either as part of a software or hardware solution taking into account that two or more functions/entities can be merged together. It is left to the implementation stage to determine the implementation stage detailed assumptions which will be discussed in more depth in the next chapter. However, the logic and flow charts of the algorithms will be described in later sections within the context of this chapter.

5.3.2 System Under Test

The System Under Test (SUT) is mainly the entity that process the requests at the core network, for this research, the core IMS components and interfaces will be considered as the main SUT. However, different architectures of IMS will be considered, and therefore different SUTs, for test and benchmarking purposes. As describes in the previous subsection, the intermediate system will apply the balance loading logic and leave it for the SUT to decide the appropriate load balancing mechanism. The topology and architecture of SUT affects the performance of the overall system. Figures 5.4, 5.5, and 5.6 show the three SUT types that will be considered in this research; the basic IMS system, two parallel IMS systems, N parallel IMS systems. Performance related aspects for different designs will be explained in a later chapter.

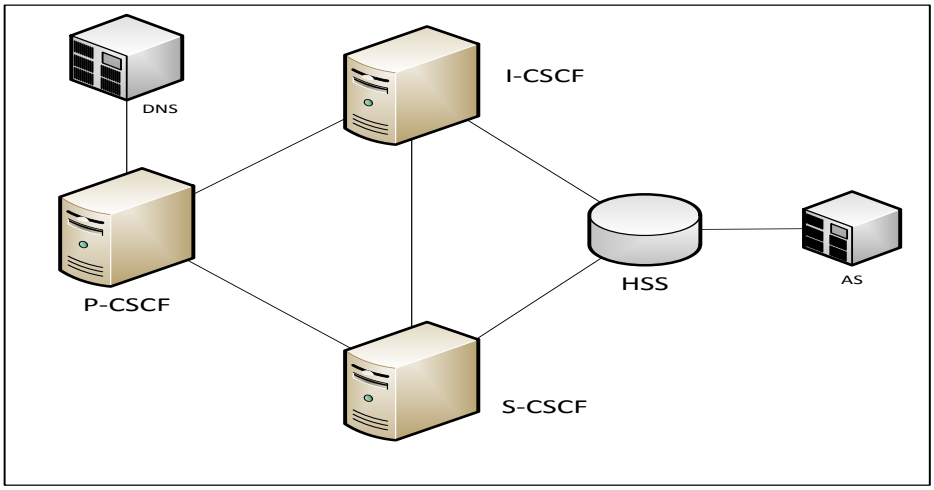


Figure 5.4 Basic IMS System

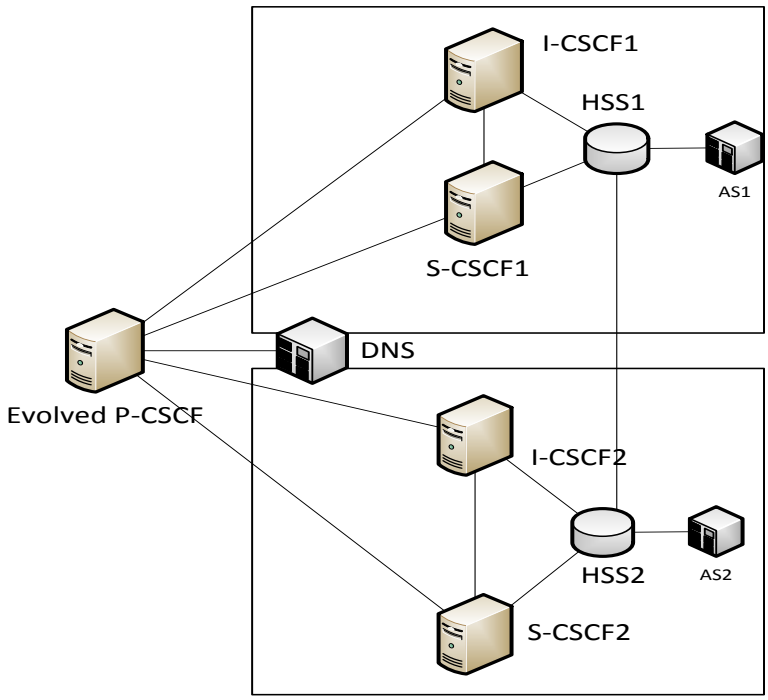


Figure 5.5 Two Parallel IMS System

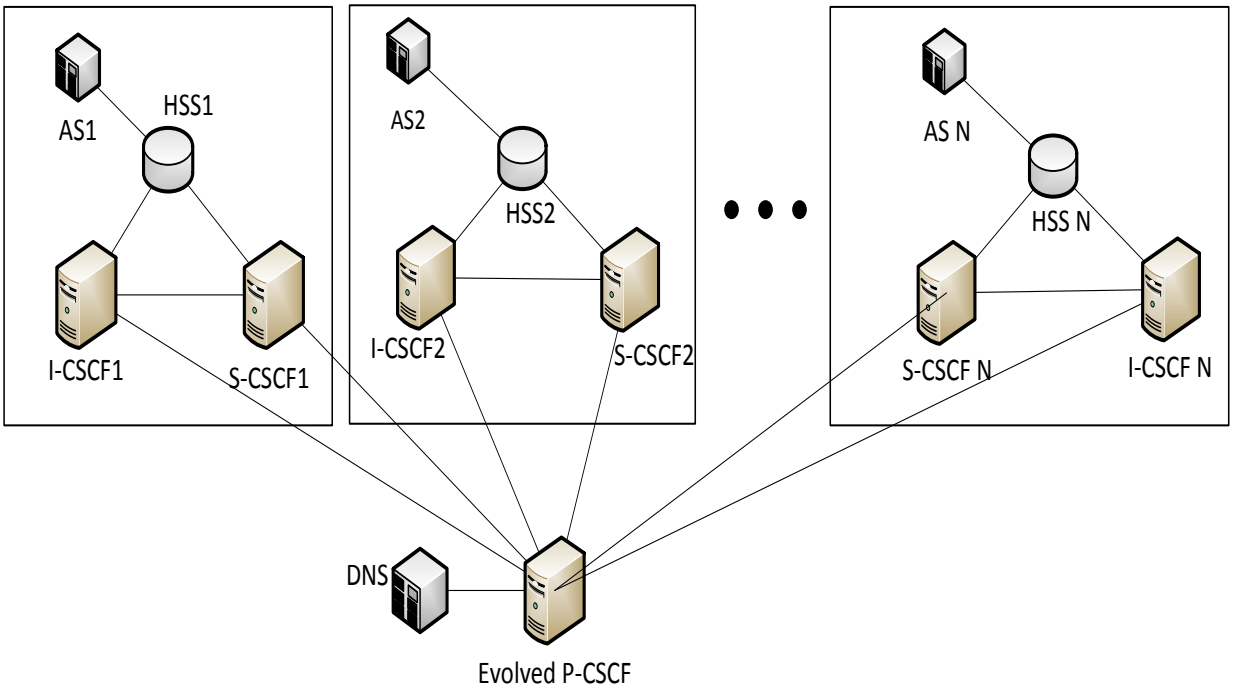


Figure 5.6 N-Parallel IMS System

It is important to note that the SUT is designed to by dynamically changing from onw setup to the other according to the load demands. Starting from one single IMS core up to N IMS cores,

this design target will allow the system to run at its maximum utilisation level while saving the processing energy and reducing the overall complexity of the system.

5.4 FUNCTIONALITY FLOWCHARTS

In this section, the performance description of the intermediate system will be presented. The flow chart of different entities is explained; figure 5.7 shows the flowcharts that describes the Forward Traffic sniffing and filtration process. The forward requests traffic log unit checks first if there is SIP traffic sensed in the input interface. Based on the logic of the design, the testing system will be generating only SIP messages, which implies that there is no need to implement a parser and a filter over the received traffic to tell whether there is SIP traffic or not. This simplifies the operation of forward traffic unit to simply get the received traffic and buffer it for a while.

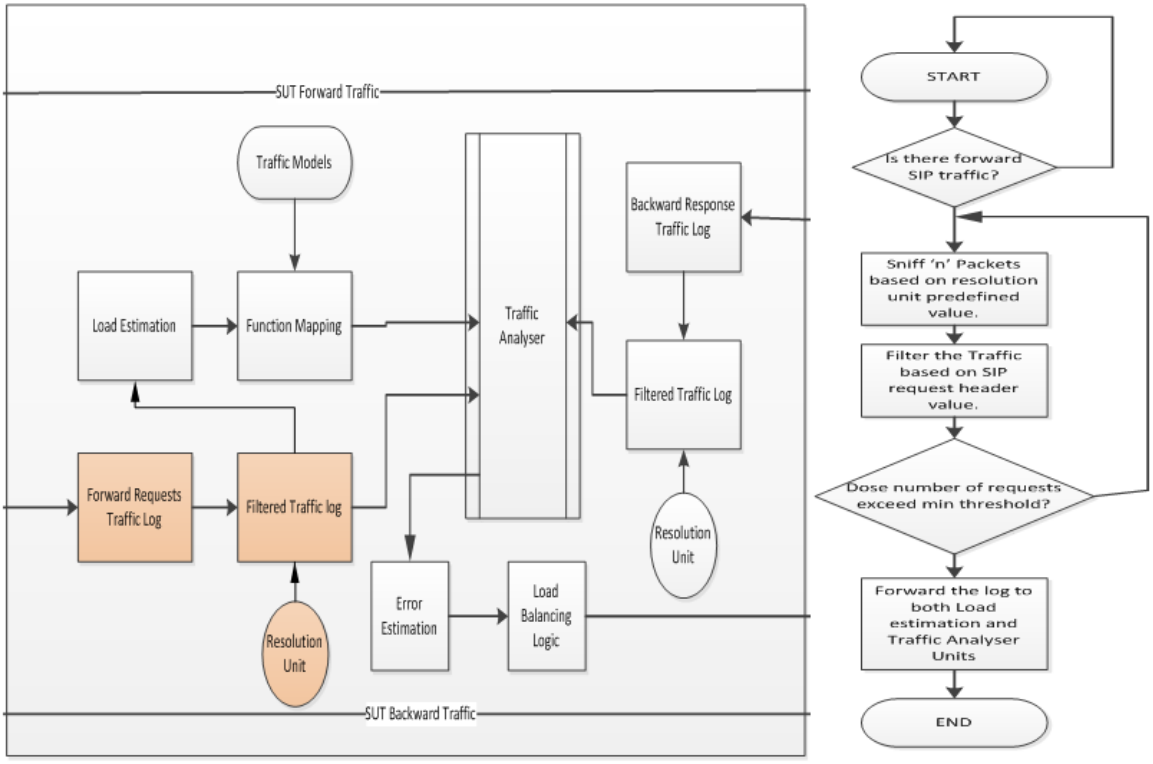


Figure 5.7 Forward Traffic Sniffing Flowchart

The number of packets need to be buffered depends on parameters set by the resolution unit, the resolution unit set a redefined or adaptively set parameters to specify the window size of the model and therefore the accuracy of performance. The traffic then is filtered based on SIP message types and compared against predefined threshold value set by the resolution centre as well. If the conditions are met, the filtered traffic is logged and forwarded to both the traffic analyser and load estimation unites for further analysis.

Following the first stage, once the load estimation get the filtered traffic log, it calculates the traffic load λ by simply counting number of request per unit time, the load may represent the registration rate if the filtered SIP traffic is based on SIP REGISTER messages header, or it could be a representation of call rate (calls/unit time) if the traffic was filtered according to

SIP INVITE header value. The load estimation is forwarded to the Function Mapping unit. Figure 5.8 shows the flowchart for load estimation and function mapping entities.

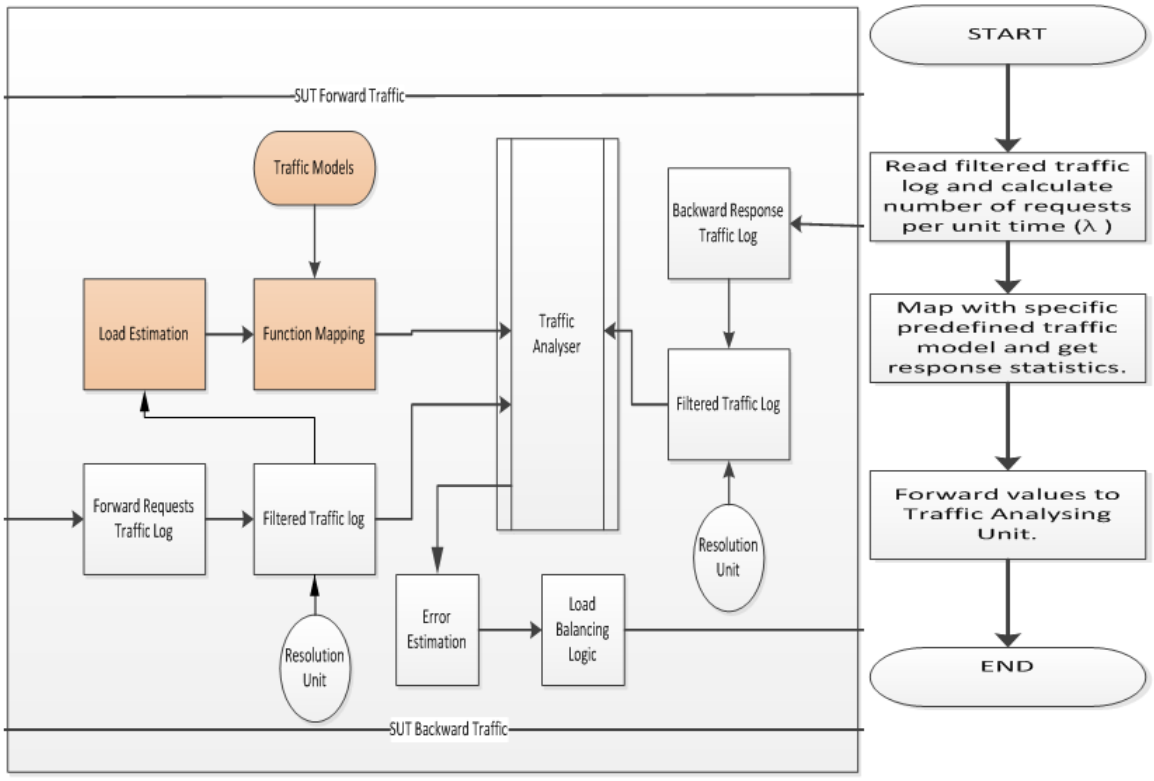


Figure 5.8 Load Estimation and Function Mapping

Depending on the load estimated value, the function mapping unit maps between the real loads estimated and predefined set of traffic models that reflect the expected stochastic performance model of the system. The stochastic model selected depends on the current SUT topology implemented (single IMS, Two parallel IMS systems, N Parallel IMS systems) and the current load generated to the SUT. Selected model parameters are then forwarded to the traffic analyser unit.

Similar to the forward traffic processing steps, the Backward Response Traffic Log Unit sniffs the traffic generated by the SUT and filter it according to the resolution unit predefined thresholds and values and filter the traffic according to the SIP header value and then forwards the logged value to the traffic analyser unit. Figure 5.9 shows the functionality flow chart of the process.

In contrary to the forward requests traffic log and filter logic, there is no need to get load estimation of backward traffic as it has no effect over the selected stochastic model. System response can be measured by another logic applied by the traffic analyser unit.

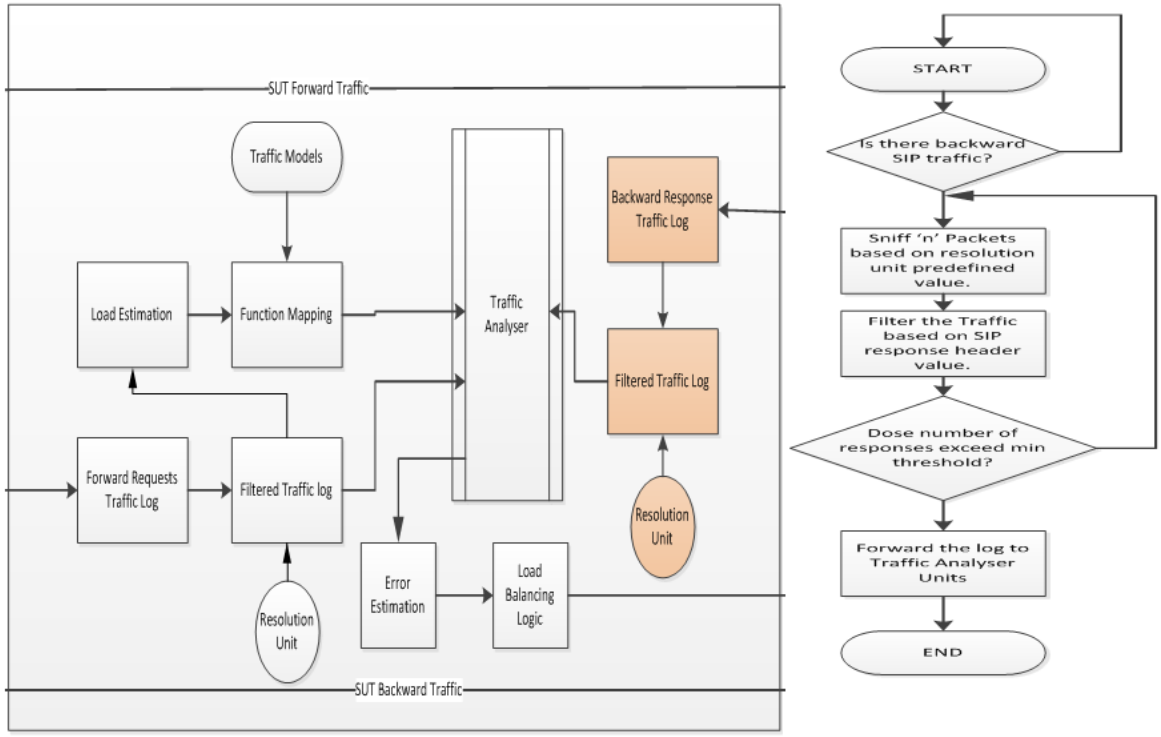


Figure 5.9 Backward Traffic Sniffing Flowchart

The traffic analyser unit get the both forward and backward traffic logs in addition to the function mapping output. It first check if the forward traffic log is generated or not, then check if the backward traffic is ready. Delays between both logs should be taken into account as it would be impossible to get both logs at the same time due to the SUT delays. If all conditions are met, the traffic analyser get both logs and buffer it for further processing.

The traffic analyser then estimate the SUT delays based on both logs records and timestamp values. The delays are then statistically analysed and compared against the peer stochastic model expected parameters. The parameters will be listed in more details in the following chapter. The aforementioned described functionality is shown in figure 5.10.

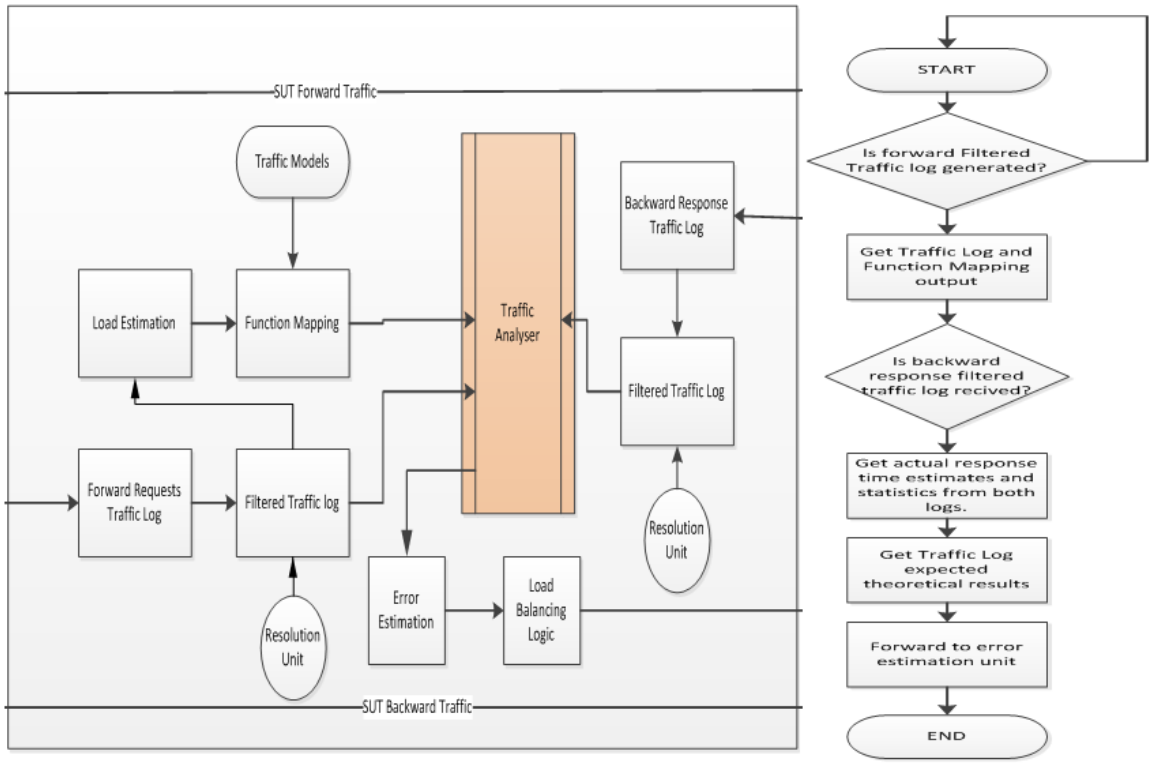


Figure 5.10 Traffic Analysis Flowchart

Both real measured statistics and model expected parameters are forwarded to the error estimation unit for further processing. Once the traffic analyser generates the output, the error estimation unit calculates the average response time of the measured responses and compare it against the theoretical waiting time expected by the model and get the error percentage value. The error is then checked against predefined threshold to reflect the accuracy of the expected behaviour of the SUT and then decided whether there is a need to apply load balancing logic or not. Figure 5.11 shows the flowchart of the described functionality.

If there is no need to apply load balancing over the SUT due to lower error values, the error estimator simply wait for another cycle of values to apply the same algorithm. The duration of the cycles is determined by the resolution unit. Otherwise, if the error exceeds the set threshold, the error estimation unit forwards the output using FeedBack Channel (FBC) to the load balancing logic which decide whether load balancing is needed and the recommended mechanism. It is left to the SUT to enforce the load balancing mechanism according to the SUT topology setup. The feedback channel input value can be displays as a four digit number as shown in table 5.1. Each entry has specific meaning and is used dynamically by the testing system to automate the test whenever needed. The value of the Feedback channel input is represented in binary notation, each value reflect specific IMS structure. The value “0011” for example, means that IMS 1 and IMS 2 are deactivated and IMS 3 and IMS 4 are activated, the testing system then will rely on the feedback channel value to decide the new destination of the generated traffic. Not all the combinations will be used for testing purposes, only those

indicated as a "BASIC" value will be considered in this study, whereas the "ADVANCED" values may be considered for a future research. The "BASIC" combinations, refers to the ones with either single active IMS core at a time (i.e. 1000 , 0100, 0010, 0001) or with consecutive trail of ones starting from the Most Significant Bit (MSB) (i.e. 1000, 1100, 1110, 1111). More will be explained later about the selection and the meaning of this specific combination.

Table 5.1 Feedback Channel Values

FeedBack Channel Input	IMS1	IMS2	IMS3	IMS4	Operation Mode
0000					Basic
1000	■				Basic
0100		■			Basic
1100	■	■			Basic
0010			■		Basic
1010	■		■		Advanced
0110		■	■		Advanced
1110	■	■	■		Basic
0001				■	Basic
1001	■			■	Advanced
0101		■		■	Advanced
1101	■	■		■	Advanced
0011			■	■	Advanced
1011	■		■	■	Advanced
0111		■	■	■	Advanced
1111	■	■	■	■	Basic

Figure 5.12 shows the flowchart along with involved entities as described before. Figure 5.12 shows the flowchart of the process described.

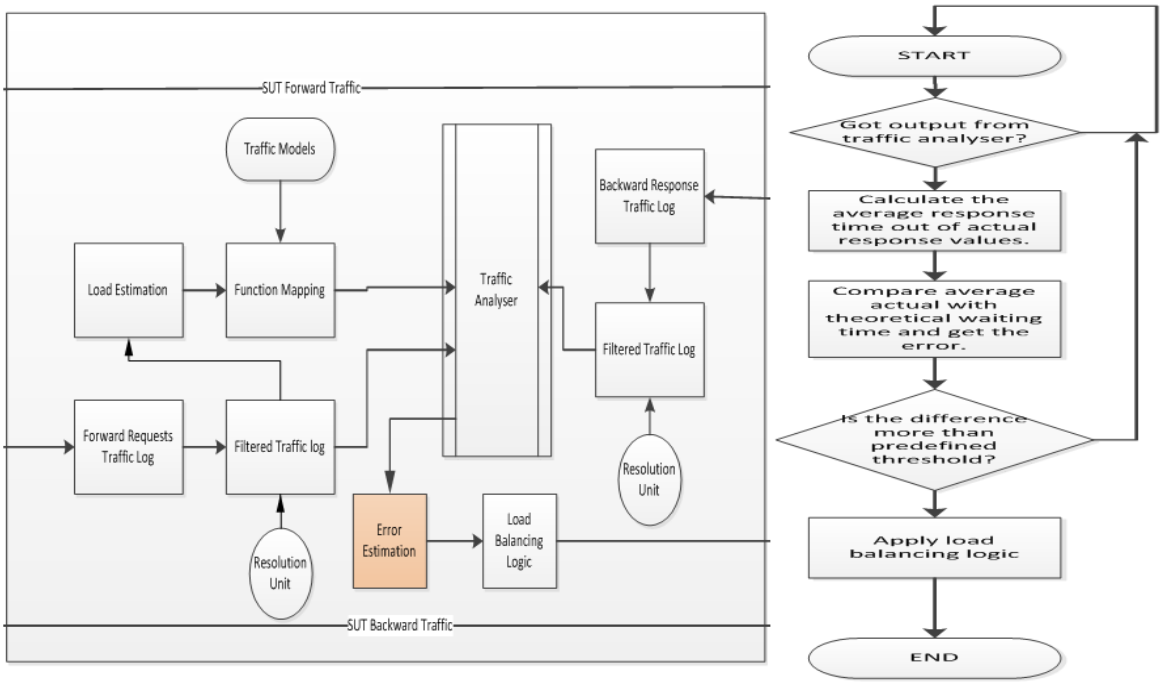


Figure 5.11 Error Estimation Flowchart

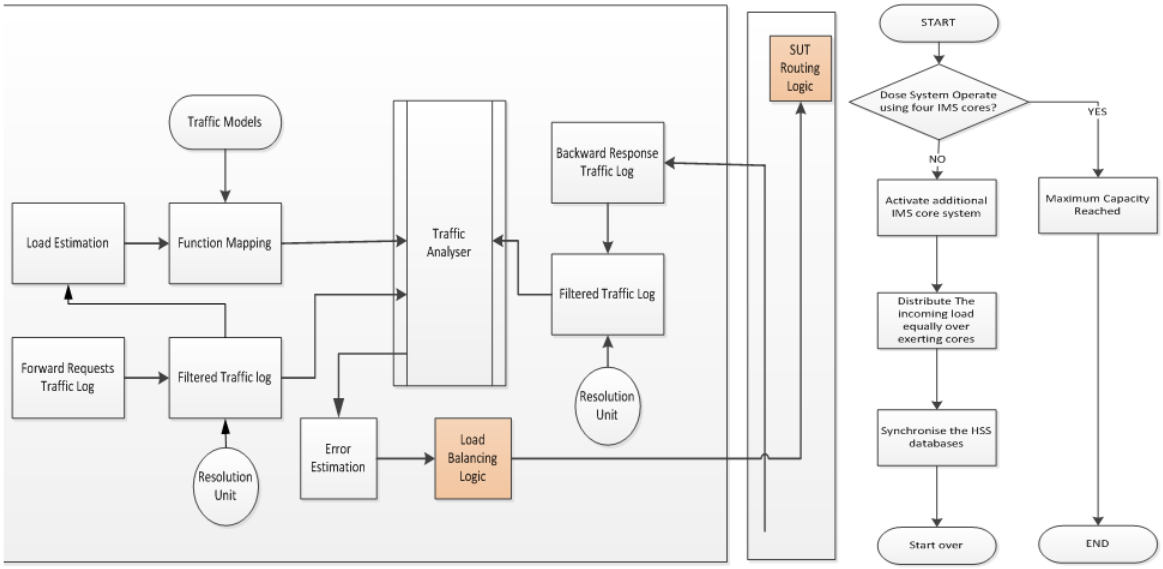


Figure 5.12 Load Balancing logic Flowchart

The overall functionality of all entities is shown in figure 5.13.

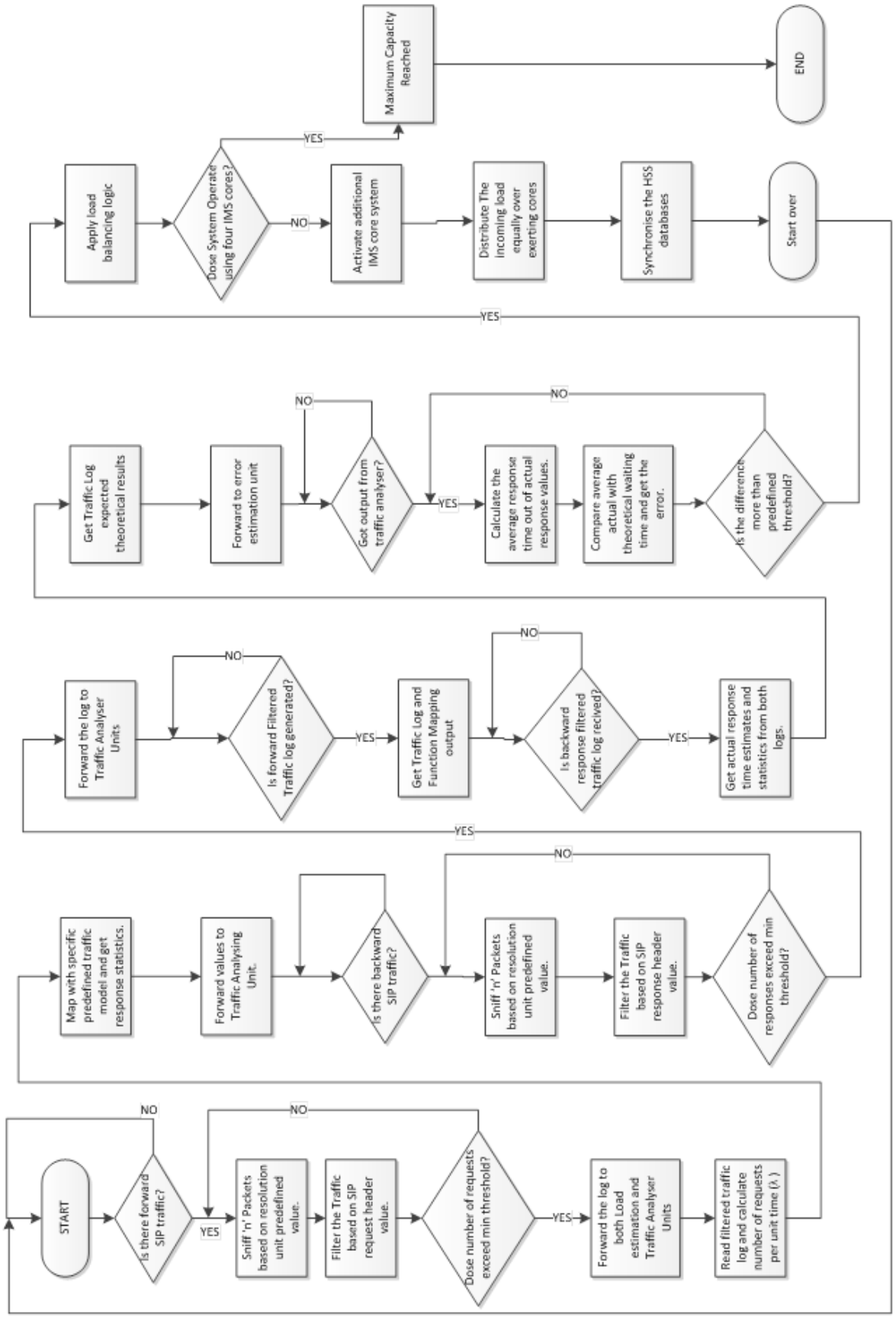


Figure 5.13 Entire System Functionality Flowchart

5.5 ALGORITHMS

The flowcharts presented in the previous section will be clarified in this section. More details are revealed in the algorithm description, this helps in understanding the functionality in more depth compared to showing the flowchart.

5.5.1 Testing System Algorithms

The testing system characterise the shape of the generated traffic by adjusting a set of parameters. Table 5.2 shows the detailed algorithm that reflects the flowchart presented in the previous section.

Table 5.2 Traffic Generation Algorithms

Algorithm 1: Traffic Generation Algorithm at the Testing system

SEND (*n, Realm, τ, FBC*, IMS1_IP, IMS2_IP,IMS3_IP, IMS4_IP)

INPUTS: Number of Users (*n*), Inter-Transmission Time (τ), Transmission Rate (λ), Domain Name (*Realm*), Feed Back Channel value (FBC), IMS1 IP address (IMS1_IP), IMS2 IP address (IMS2_IP), IMS3 IP address (IMS3_IP), IMS4 IP address (IMS4_IP).

OUTPUTS: Generate Real-time IMS Registration Traffic.

1: if (*current user index*) < *n*
2: For Each User *i*
3: Get *FBC* updated value
4: If ((*FBC*) EQUAL TO (1000))
5: Send Register Message to IMS1 using *Realm* as a domain name and IMS1_IP as a Destination Address
6: Wait for *200OK* Reply message
7: Pause (τ)
8: Else if ((*FBC*) EQUAL TO (1100))
9: Send Register Message to IMS1 and IMS2 using *Realm* as a domain name and IMS1_IP and IMS2_IP as a Destination Addresses
10: Wait for *200OK* Reply message
11: Pause (τ)
12: Else if ((*FBC*) EQUAL TO (1110))
13: Send Register Message to IMS1, IMS2 and IMS3 using *Realm* as a domain name and IMS1_IP, IMS2_IP and IMS3_IP as a Destination Addresses
14: Wait for *200OK* Reply message
15: Pause (τ)
16: Else if ((*FBC*) EQUAL TO (1111))
17: Send Register Message to IMS1, IMS2, IMS3 and IMS4 using *Realm* as a domain name and IMS1_IP, IMS2_IP, IMS3_IP and IMS4_IP as a Destination Addresses.
18: Wait for *200OK* Reply message
19: Pause (τ)
20: Else
21: END Transmission and Record Total number of Transmissions.
22: Else
23: END Transmission as Total Number of Users are Registered

5.5.2 Intermediate System Algorithms

The intermediate system algorithms are listed in the same order the flowcharts was presented in the previous section. Table 5.3, 5.4, 5.5, 5.6, and 5.7 show the algorithms for he flowcharts

presented in the previous section describing the intermediate system functionality in the same order.

Table 5.3 SUT parsing Traffic Algorithm

Algorithm 2: SUT Forward Traffic Parsing Algorithm at the Intermediate system

ReceiveRequests (*n, Realm*)
INPUTS: Number of Received Requests per iteration (*n*) , Domain Name (*Realm*),
OUTPUTS: Requests Rate (λ)

1: Check the resolution unit value (n)
2: if (no. of Received Packets < *n*)
3: *Receive* additional packets for specified Realm
4: *Parse* the header of the message
5: *Check* if there is additional IMS Requests
6: *Update* the (no. of Received Packets) counter
7: else
8: Count number of Requests and record the reception time for each one.
9: Calculate the Requests Flow Rate (λ).

Table 5.4 Traffic Model Mapping

Algorithm 3 : SUT Forward Traffic Model Mapping Algorithm at the Intermediate system

MappingModel (λ,*Realm*, *FBC*)
INPUTS: Requests per Unit Time (λ), Domain Name (*Realm), FBC value*
OUTPUTS: Average Expected Theoretical Delay (γ)

1: Check the Load Estimation unit value (λ)
2: Check the FBC value.
3: Use λ and FBC to get approximation model of the system
4: Calculate the Theoretically Expected Average Delay of the Model.
5: Forward the Values to the Traffic Analyser Unit.
6: if (FBC=(1111))
7: DO NOTHING
8: END

Table 5.5 SUT Backward Traffic Parsing Algorithm

Algorithm 4: SUT Backward Traffic Parsing Algorithm at the Intermediate system

ReceiveReplies (*n, Realm*)
INPUTS: Number of Received Replies per iteration (*n*) , Domain Name (*Realm*),
OUTPUTS: Replies log and timestamps

1: Check the resolution unit value (n)
2: if (no. of Received Packets < *n*)
3: *Receive* additional packets for specified Realm
4: *Parse* the header of the message
5: *Check* if there is additional IMS Replies.
6: *Update* the (no. of Received Packets) counter
7: else
8: Count number of Replies and record the reception time for each one.

Table 5.6 Traffic Analyser Algorithm

Algorithm 5: Traffic Analyser Algorithm at the Intermediate system

Analyse (*n*, *Realm***,** γ, BRTS, FRTS)

INPUTS: Number of Users (*n*) , Average Expected Theoretical Delay (γ), Domain Name (*Realm*), Backward Replies log and timestamps (BRTS), Forward Requests log and timestamps (FRTS)

OUTPUTS: Error Threshold (α)

1: For (Each Iteration)
Compare BRTS and FRTS Values and find the difference between them.
2: For Each User i < n, get the total response time.
3: Calculate the Average total response time (TAVG) for all n users.
4: If (TAVG - γ) > α
5: Activate Load Balancing Mechanism by setting flag (LB="1")
6: else
7: DO NOTHING
8: END
9: START NEW ITERATION

Table 5.7 Load Balancing Algorithm

Algorithm 6: Load Balancing Algorithm at the Intermediate system

Analyse (*Realm***,** LB, IMS1_IP, IMS2_IP,IMS3_IP, IMS4_IP, FBC)

INPUTS: Domain Name (*Realm*), Load Balancing Flag (LB), IMS1 IP address (IMS1_IP), IMS2 IP address (IMS2_IP), IMS3 IP address (IMS3_IP), IMS4 IP address (IMS4_IP), current Feed Back Channel Flags (FBC).

OUTPUTS: Set New Feed Back Channel Flags (FBC).

1: For (Each Iteration)
If LB = "1"
2: Get current FBC flags state
3: if FBC = "1000"
4: New_FBC = FBC Logic OR "1100"
5: Send New_FBC to the Testing System for Next Transmissions.
6: else if FBC = "1100"
7: New_FBC = FBC Logic OR "1110"
8: Send New_FBC to the Testing System for Next Transmissions
9: else if FBC = "1110"
10: New_FBC = FBC Logic OR "1111"
11: Send New_FBC to the Testing System for Next Transmissions
12: else if FBC = "1111"
13: DO NOTHING
14: END
15: START NEW ITERATION
16: END

5.6 SUMMARY

In this chapter, the framework design methodology was described in a systematic way. The framework entities was identified and explained. A precise functionality description is crucial to set the evaluation baselines in the following steps. The system performance was identified by a set of flowcharts and algorithms that will be referred to in the later chapters whenever a specific scenario is implemented. The different SUT implementations are described, different implementations create differences in performances for different scenarios.

The proposed novel framework has introduced a new mechanism for IMS systems via merging multiple IMS subsystems together and having a feedback channel for reporting load conditions to the intermediate system continuously. The value of the feedback signal indicates the currently available and busy IMS cores, this allows the system to manage the load according to the current conditions and the predicted load.

The algorithms presented describes in detail the functionality of the IMS components. It was shown that the framework was designed to operate in two modes of operation; open-loop and closed-loop modes. Moreover, the system has the ability to adapt to the load introduced via configuring the SUT dynamically by activating additional IMS core component to absorb the overshoot of added traffic during disaster scenarios.

Chapter 6: PERFORMANCE EVALUATION AND ANALYSIS OF THE ENHANCED IMS FRAMEWORK

6.1 INTRODUCTION

In this chapter the performance of the overall system will be carried out according to the previously suggested algorithms and framework design. A testbed that reflect the high level framework model will be explained. A set of scenarios that shows the strength and weaknesses of the proposed solution will be defined and conducted. And finally the results will be discussed and analysed. The chapter will be organised as shown in figure 6.1.

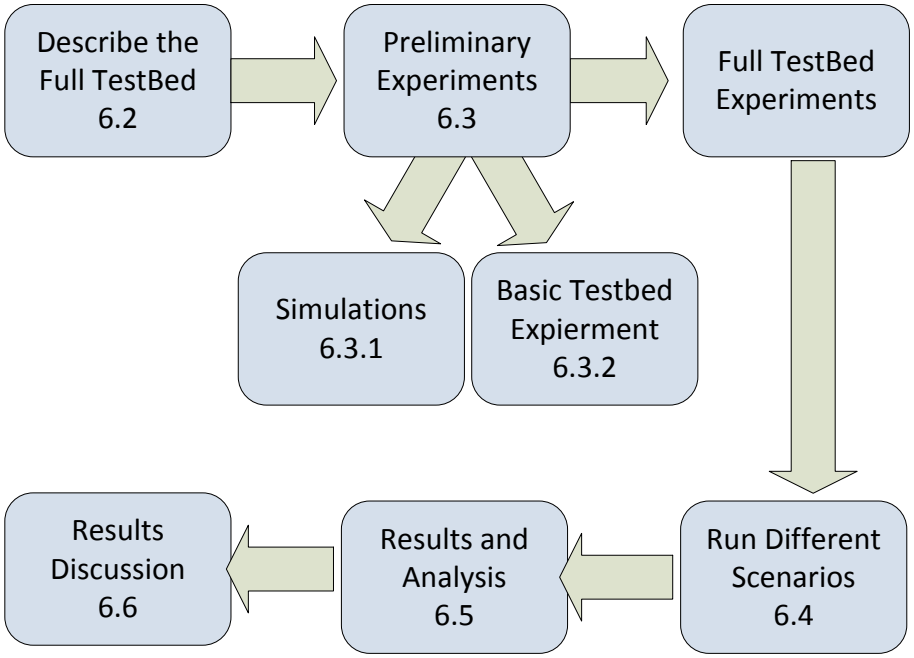


Figure 6.1 Chapter Organisation

6.2 TESTBED SETUP

The testbed is designed to validate the proposed framework in the previous chapter. It is important to mention that the logical entities of the operation logic that represent the three main entities of the framework (Testing, Intermediate, and SUT) may be merged or separated into one or more physical entity with no effect over the intended performance and functionality

design goal. Figure 6.2 shows the testbed implemented in the Wireless Systems Lab.

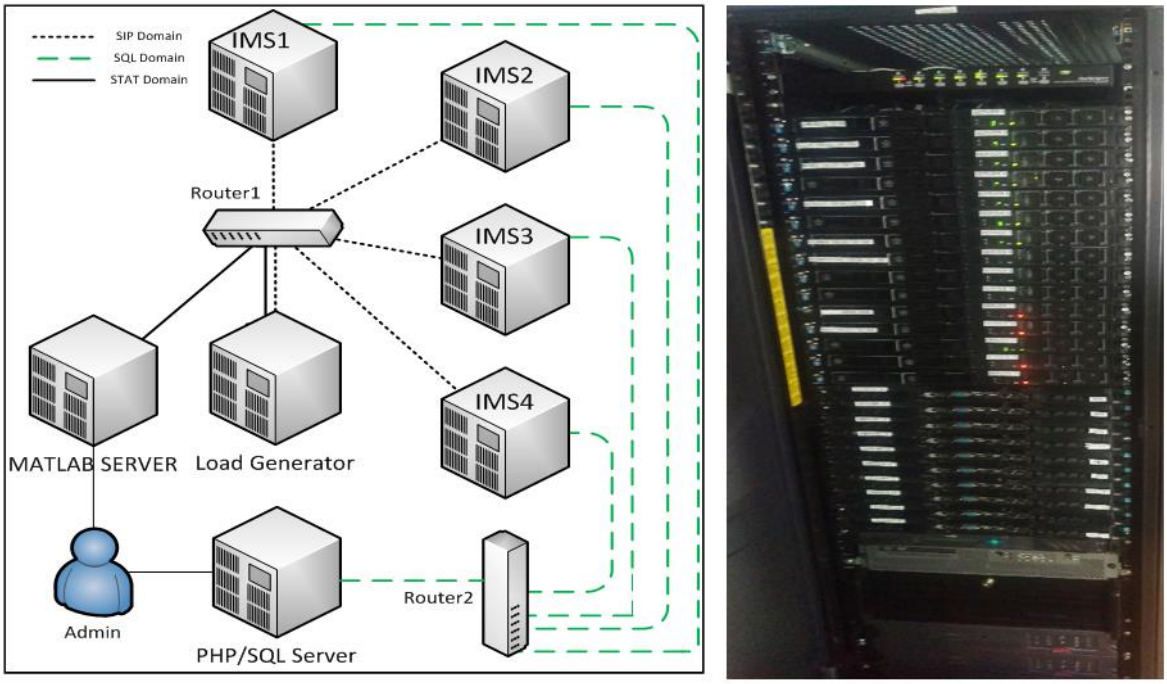


Figure 6.2 Testbed setup

6.2.1 Load Generator

The testbed represents mainly the core network abstract, the aggregation traffic of multiple access technologies is modelled by the load generator and introduced then to the other core network entities.

The main design goals of the load generator can be summarised as follows:

1) To model a traffic load that reflect different realistic scenarios.
2) To work independently of the other core network entities.
3) To deploy the load balancing functions.
4) To have part of the PCSCF functions.

To ensure satisfying the aforementioned design goals along with maintaining implementation simplicity; the Load Generator is split into two logical entities; the Traffic Generator and the Load Balancing Unit. As the load balancing operation needs a feedback information that reflect the current conditions of the system, it was crucial to have such logical separation between the generation and balancing logics to have a more realistic mode of operation.

6.2.2 MATLAB Server

The MATLAB Server is responsible for analysing the traffic data and change the mode of operation accordingly. It gets the data from both the load generator unit, the SUT unit and the optional human interface to determine the current mode of operation and the current system state and when to trigger the system to change to the next state according to certain set of measures. The main design goals of the server is summarised as follows:

1) Not to intervene with the normal operation and traffic flow between the Testing System and the SUT.

2) Running in Real-time or near Real-time mode of operation.

6.2.3 IMS Servers

The IMS servers represent the modelled SUT. Each one is composed of a set of servers that implement the basic functionality expected for IMS with the exception of running a dedicated DNS server for each one and migrating part of the PCSCF functions to the Load Generator Server which is referred as Evolved-PCSCF.

The IMS core is based on OpenIMSCore (Fokus, 2004) which is open source implementation developed by FOKUS Institute for Open Communication System, though it was first proposed in 2004, it is still considered valid for our testing experiments as the same basic IMS functionality is still deployed nowadays. The server embed the IMS Call Session Control Functions CSCFs; such as PCSCF, I-CSCF, and S-CSCF. In addition to the home Subscriber Station HSS, which are all considered the core architecture for the next generation networks that was specified by 3GPP. The purpose of the experiment is to test the capacity of the IMS in terms of the maximum number of users that can be adapted by the system without any stability issues.

6.2.4 PHP/SQL Server

The SQL server is implemented to monitor the HSS databases for each IMS server and to insure that the databases are consistent and updated automatically whenever there is a change detected in one of the entries. The distributed database approach was the selected and preferred option compared to a single centralised database in the core network.

6.3 PRELIMINARY ANALYSIS

This section will present the evaluations and experiments made to identify the gap in the research field. It tests the IMS system under different multimedia signalling loads to tell if it is good enough to be considered for mission critical applications. The proposed enhanced IMS model evaluation will be presented in another section. All the findings in this section will be considered as a benchmark or a reference design model to better understand the signalling performance. Starting with model simulations results to test the SIP signalling overhead and performance aspects, then the basic experimental testbed model that will show the performance evaluation of the SIP and IMS signalling overhead.

6.3.1 Simulation Experiments

To facilitate investigation of the LTE system, especially the SIP signalling performance over LTE communication network, the OPNET simulator was used to create a scenario with multiple users initiating calls. This enabled the SIP performance metrics to be used to measure the efficiency of the system and its capacity tolerance. Figure 6.3 shows the created setup.

Simulation Setup and Scenarios

In this research study, OPNET Modeller provides the required level of simulation capabilities to implement and model different multimedia applications over LTE. The system design that was implemented and investigated is shown in figure 6.3 and is based on the configuration parameters shown in Table 6.1. The implementation of the LTE network system is based on a single Evolved Packet Core (EPC) that serves two eNBs, each with four clients. The clients in eNB1 make SIP-based VoIP calls to the clients in eNB2 through the EPC in a Normal distribution call generation system, using a fixed length calls. The EPC is then connected to the SIP server, which reflects the performance of the P-CSCF in the IMS that manage the registration, call initiation and call termination processes using the SIP signalling system using the IP cloud. In this research, the simulations were performed without any background traffic in the LTE system and the IP network. This enables us to study the actual performance level for SIP-based VoIP applications within a best effort environment which helps with the results accuracy. It should be noted that this research has not considered any clients mobility performance implication over signalling delays, it is left as a future work to discuss it further.

The simulation implementations has considered four scenarios based on the design shown in figure 6.3 and the simulation parameters shown in Table 6.2. The first scenario represents the basic implementation for VoIP applications over LTE using a single pair of UEs between client A-1 in eNB 1 and client B-1 in eNB 2. This scenario examines the best-case implementation of the assigned network system with only one single call at a time. The second scenario has an additional connection with multiple calls with another pair of UEs (client A-2 and client B-2) added to the first scenario. A similar thing is true of the third scenario, where additional pairs of calls are added (client A-3 and client B-3). Finally, the fourth scenario has yet another additional pair between client A-4 and B-4. This gradual increase in the pairs of SIP-based VoIP calls allowed the performance of the SIP signalling system over LTE based communications with additional VoIP calls between different clients to be checked. The highest load of VoIP calls is represented in the fourth scenario that consumes higher bandwidth over LTE where all clients in each eNB are calling one single client in the other eNB. Therefore, the results of these implemented scenarios can be compared and studied throughout the research study in terms of the performance for SIP signalling and efficiency for LTE system.

Table 6.1 Simulation Parameters in OPNET

A. LTE Network System			
Number of Simulations	4	**Simulation Seed Number**	128
Simulation Duration:		30 Minutes = 1800 Seconds	
Number of EPC:	1	**Background Traffic**	0%
Number of eNB:	2	**Number of nodes for each eNB:**	4
Antenna Gain for eNB:	15dBi	**eNB Maximum Transmission Power:**	0.5 W
eNB Receiver Sensitivity:	- 200 dBm	**eNB Selection Threshold:**	- 110 dBm
B. Applications: SIP Based VoIP			
VoIP Calls (Unlimited)	**Call Duration** 10 Sec	**Caller** Node A	**Callee** Node B
Maximum Simultaneous Calls	**SIP Server** Unlimited Call/ Second	**User Agent (Caller/Callee)** 1 call at time between each pair	**Voice Codec:** **GSM** 13 Kbps
Calls Start Time Offset:		Normal (150 sec, 100 sec)	
Calls Inter-repetition Time:		Normal (20 sec, 5 sec)	

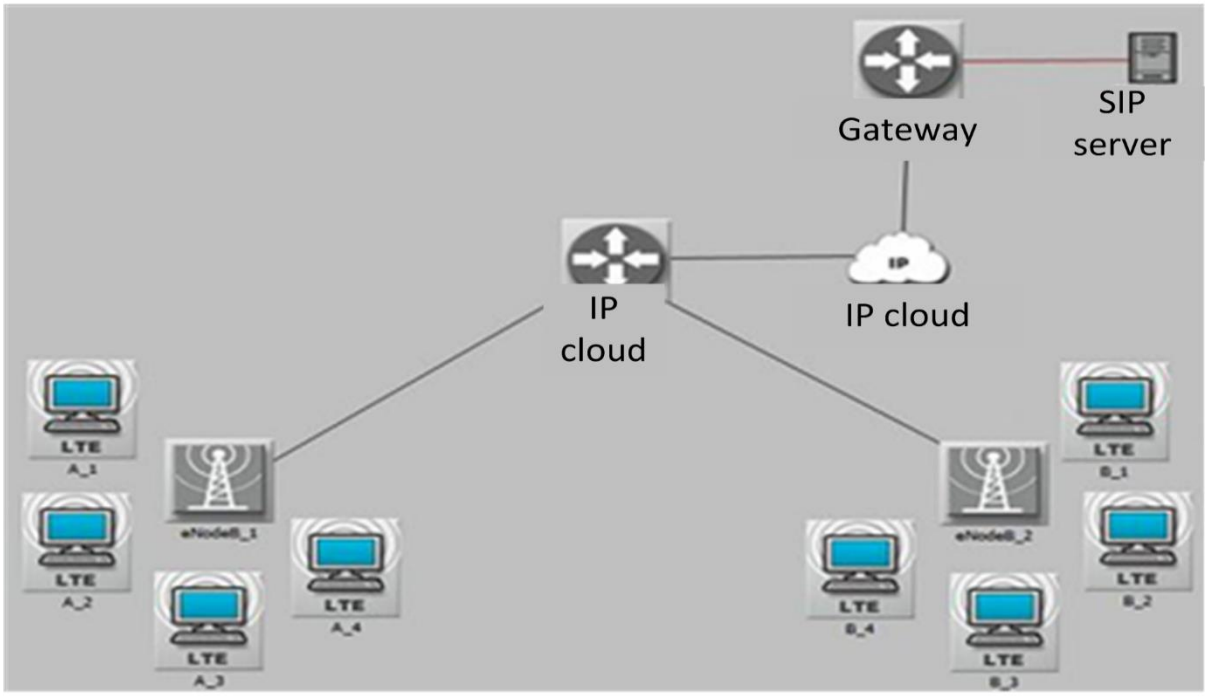


Figure 6.3 System design and implementation for SIP-based VoIP applications over LTE network system in OPNET

6.3.2 Simulation Results

As the main considerations in this study are SIP signalling and LTE performance for mission critical systems, the results focus on the call setup time and the related LTE performance metrics. The optimum number of initiated calls for each pair of calls falls between 150 and 180 for 30 minutes of simulation time with a uniform based distribution system for calls initiation. Table 6.2 shows the number of rejected calls in the overall system for the four scenarios with the implemented normal based system. The number of rejected calls has increased with the increased number of initiated call pairs. For calls implemented from Caller A-1, the number of failed calls *initiation* processes had been increased with the increased number of call pairs with scenarios S2, S3, and S4, where the total initiated calls over all scenarios is 56. This increased fail rate during the call initiation stage is mainly related to the inferior processing performance of the SIP servers' and LTE system performance.

Table 6.2 Calls Statistics from Simulation Results

SIP calls statistics for the Implemented Scenarios				
Scenario	**S1:** 1Pair	**S2:** 2Pairs	**S3:** 3Pairs	**S4:** 4Pairs
Number of Calls Rejected in the overall system	45	95	152	218
Number of Calls Initiated from	56	56	56	56

Caller A-1				
Number of failed calls initiation for calls from **Caller A-1**	27	30	38	34

Call Setup Performance

The purpose of studying the call setup time is to facilitate analysis of the SIP signalling performance during the main SIP signalling stage over different call sessions. As long as the call setup time for the majority of initiated SIP-based calls were in the acceptable range, the performance of the SIP signalling system falls in its acceptable level (D. Malas 2011) (Voznak and Rozhon 2010). Fig. 6.4 represents the average call setup time for all successful VoIP calls for the four implemented scenarios.

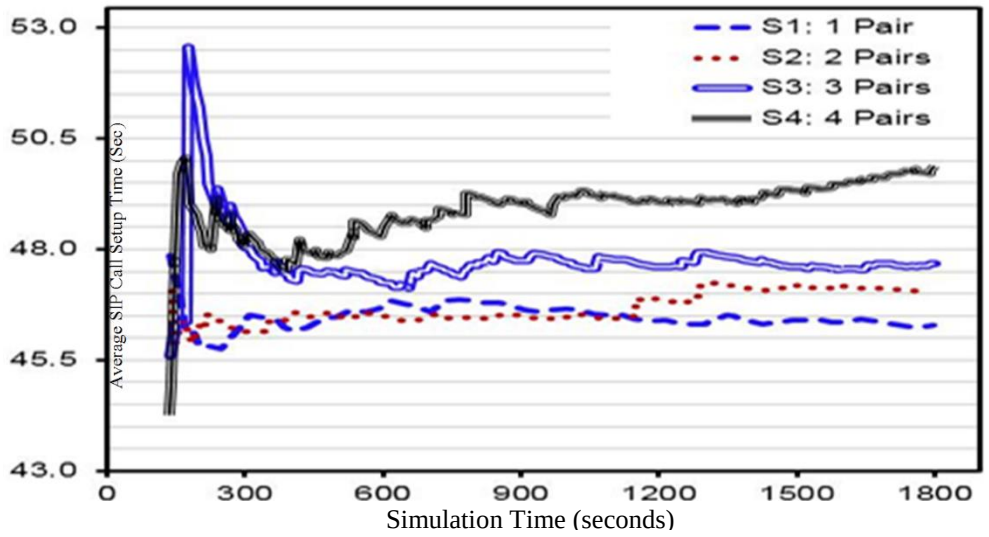


Figure 6.4 Call setup Delay

The results show that the scenario with only one pair of calls had the lowest average call setup with times ranging from between 46 and 47 ms, and increased up to 48.5 ms when the scenario involved two pairs. With three pairs of VoIP calls, the average call setup time increased from 47 ms to 49.5 ms. The longest call setup time registered was for the fourth scenario in which four pairs of VoIP connections were active at the same time. These simultaneous calls affected the SIP signalling performance and increased the average delay by up to 50.5 ms. In general, the call setup time for successfully initiated calls over all scenarios is still at an acceptable level when considering the performance of the SIP signalling system. This was due to implementing the LTE network system without having extra overloads due to added background traffic.

LTE Downlink Packets Dropped

The LTE parameters of the implemented system have a direct effect on the performance of the running applications. Real-time applications can be enhanced if the LTE system performance behaviour has been considered. The average number of packets dropped starts between 1 and 3 packets/sec for the single pair scenario and increases to between 6 and 18 packets/sec for the four pairs scenario as shown in Figure 6.5. The downlink packets dropped of LTE system has a direct link to successful rate of the SIP sessions in which it is directly proportionally increasing.

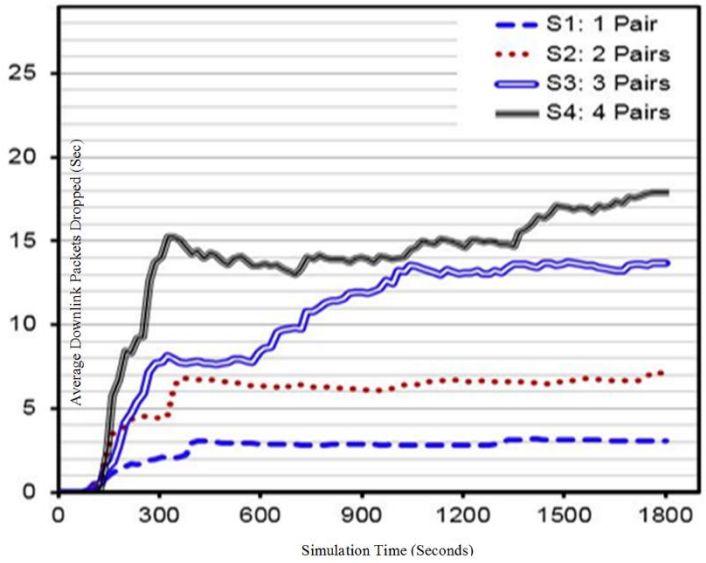


Figure 6.5 Average Packets Dropped

The LTE system delays in the transferred data between LTE components affect the performance for real-time applications. The average LTE delays with one and two pairs of VoIP calls is between 2 ms and 2.7 ms, as shown in Figure 6.6. The average LTE delays for three pairs of VoIP calls is from 2.4 ms to 3.5 ms, and between 2.5 ms and 4.3 ms with four pairs of calls. The longest delays mostly occur at the system start-up time and stabilise later during the simulation time.

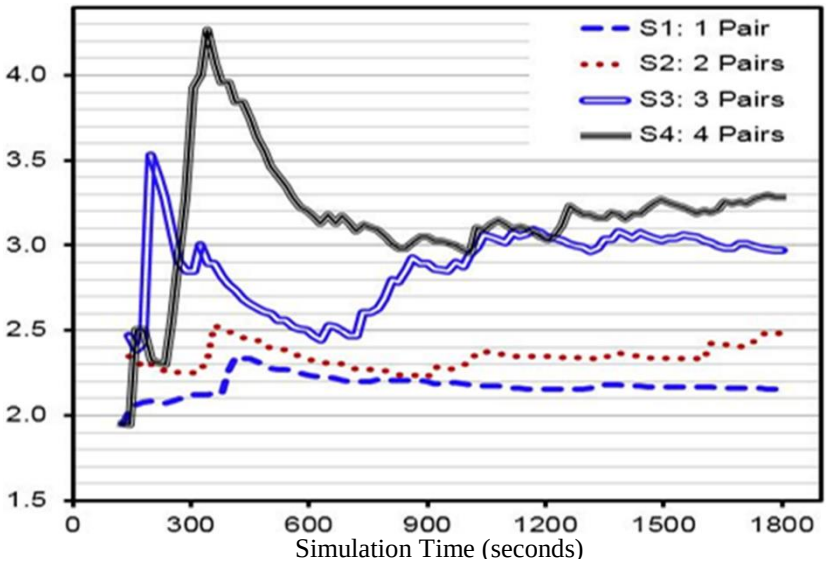


Figure 6.6 Average LTE Delays in second for Caller A-1 node

Simulation Results Summary

Based on the testbed results, it has been shown that the scalability of the system is negatively affected by the increasing number of registration requests sent to the system. It was found the RRps increases exponentially when the number of users is increased (up to a limit 1,100 users) before the RRps starts to decrease leading eventually to system degradation and failure. This is clearly shown by RRps for both scenarios 6 and 7, where the RRps of scenario 7 is much less than the RRps for scenario 6 (although the number of users increased by 200 users).

Based on the simulation results, it is clear that there is increasing delay in the call setup time when the LTE communication system is used. This delay increases as the number of served client also increases, which indicates that the Delay requirement or the maximum number of users that can be served at a time may not meet the mission critical service requirements. Hence, the need for decreasing the gap of call setup delay for commercial broadband systems compared with other dedicated mission-critical communications systems is of great importance and considered one of the main challenges for mission-critical communications. This means that there is a need for a new mechanism that minimises access delay overhead by exploring the LTE and IMS domains in addition to the interfaces between LTE and IMS and the interface between LTE and User Element.

The simulation has been implemented using static mobile nodes (that is, the positions of the nodes are fixed). Hence, there is no handoff added complexity for the nodes moving between two cell domains. If mobility were to be considered (that does not imply moving nodes only but rather a dynamic topology) then support for handoff mechanisms between the subscriber stations and different base stations would need to be considered. Therefore, further testing of different communication scenarios for an end-to-end connectivity over LTE communication system is needed. For such dynamic topology, the need for measuring the overall performance of the system in terms of SIP signalling and data streaming delay is crucial.

6.3.2 Basic Testbed Experiment

6.3.2.1 Basic Experiment Setup

Figure 6.7 shows the experimental topology of the testbed. The testbed is composed of four main parts: the Packet Generator; IMS core which is based on Open-IMS-Core model (Fokus 2004); Packet Analyser, and Domain Name Server (DNS). The parts function and operation are as follows:

- Packet Generator: The packet generator is responsible for simulating virtual clients that then generate concurrent calls that are transmitted in a serial or parallel manner by a theoretically unlimited number of users, figure 6.8 shows the basic traffic generator used for preliminary experiments set. Due to the focus on SIP and IMS performance, the Packet Generator is designed to send SIP Register Message (as defined by RFC 3261) in addition to SIP invite and bye messages. All the messages are transported using the UDP where the sender port address is dynamically allocated so as to avoid using restricted ports at the sender or server sides. The GUI interface of the Packet Generator enables the user to select a predefined set of users and the Proxy IP address. Finally, the SIP request-sending pattern is selected to be either serial or parallel.

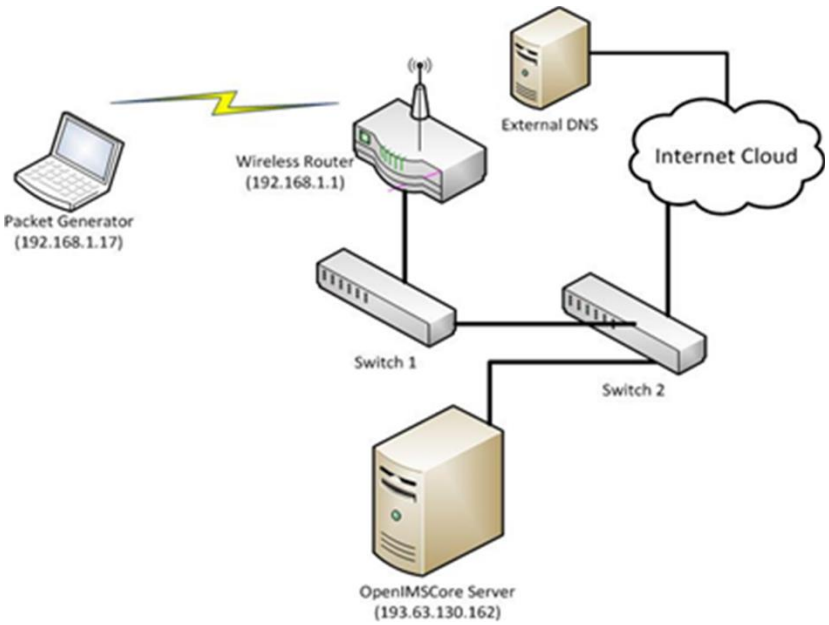


Figure 6.7 Experiment Testbed

- IMS Network: The IMS core is based on Open-IMS-Core (Fokus, 2004) developed by the FOKUS Institute for Open Communication System. The server embed the IMS Call Session Control Functions CSCFs; such as PCSCF, I-CSCF, and S-CSCF, in addition

to the home Subscriber Station HSS. All are considered part of the core architecture for the next generation of networks as specified by 3GPP. The purpose of the experiment is to test the capacity of the IMS in terms of the maximum number of users that can be adopted by the system without causing stability issues.

- Packet Analyser: The packets sent by the Packet Generator are monitored using Wireshark as a packet analyser at the sender side. The trace files extracted from the packet analyser help in calculating the KPI values for both SIP and IMS.
- DNS: an external Domain Name Server (DNS) to resolve the IP addresses of all servers in the system setup.

Based on the previous setup the experiment aimed to evaluate the SIP performance over the IMS using either a wired or a wireless connectivity with the server. For this purpose, the register message delay was calculated by running Wireshark at the packet generator side and calculating the difference between the sent registration request time and the 200OK response reception time. The data was then exported using MATLAB and analysed the Probability Density Function (PDF) and Cumulative Density Function (CDF) curves calculated. These provide a better understanding of the variance in Registration delay within the same scenario and among different running scenarios.

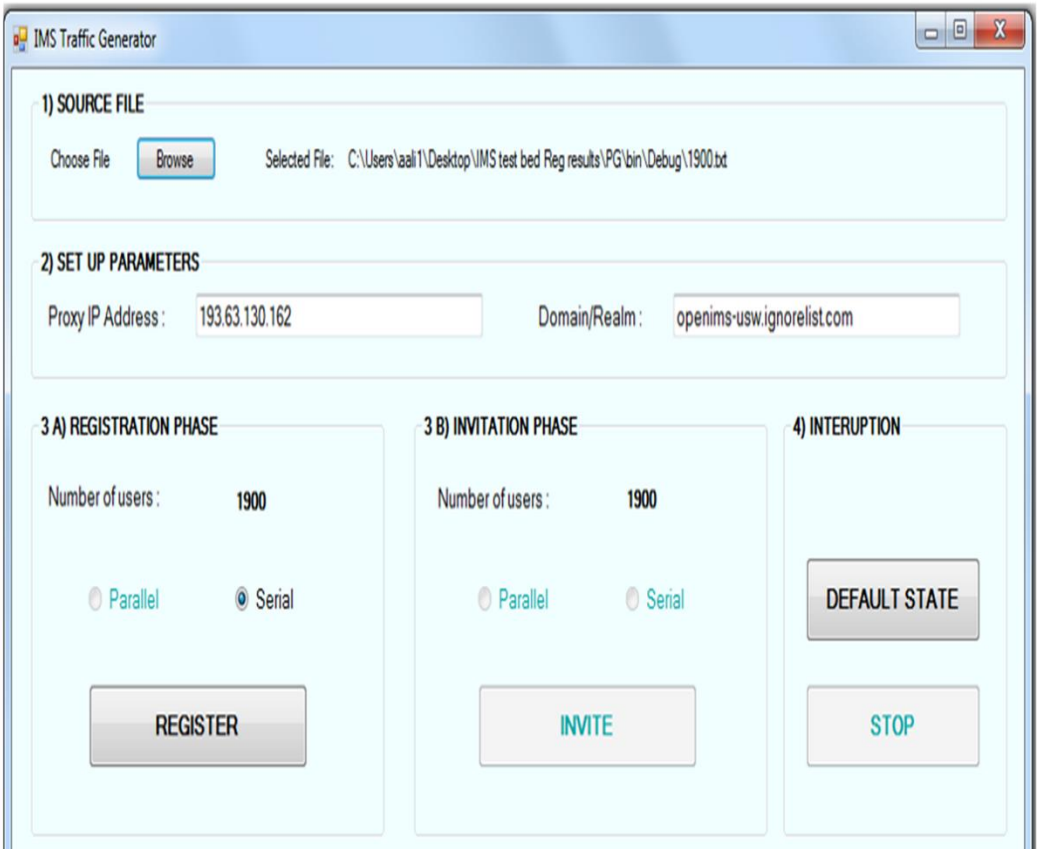


Figure 6.8 Packet Generator GUI

Figure 6.8 shows the GUI interface for the Packet Generator. The user first selects a predefined set of users' databases and the Proxy IP address (which is the IMS server IP). In addition, the domain name is inserted and the SIP request sending pattern selected (either series or parallel).

Pressing REGISTER initiates the sending of the registration requests (one per user) consecutively and dynamically. The packets sent by the Packet Generator are monitored using Wireshark at the sender side while the log screen at the packet generator records the time stamp of the sent requests and received responses, which helps in calculating the end-to-end application delay (the time between peer application layers).

Based on the previous setup, the experiment aimed to evaluate the SIP performance over the IMS using either wired or wireless connectivity to the server. For this purpose, the register message delay is calculated by running Wireshark at the packet generator side and calculating the difference between the sent registration request time and the 200OK response reception time. The data is then, using MATLAB, exported and manipulated in order to generate curves for a better indication of the variance in Registration delay with time. The experiment was repeated multiple times, each time the number of users was incremented in both the wired and wireless scenarios.

6.3.2.2 Experiment Results

In this scenario, the packet generator was wired directly to the router and the number of users sending the registration request were incremented in steps of 200 in the range from 100 to 1,300 users. Figure 6.9 shows the PDF and CDF of the registration delay for 100 users while Figure 6.10 shows the PDF and CDF for 500 users.

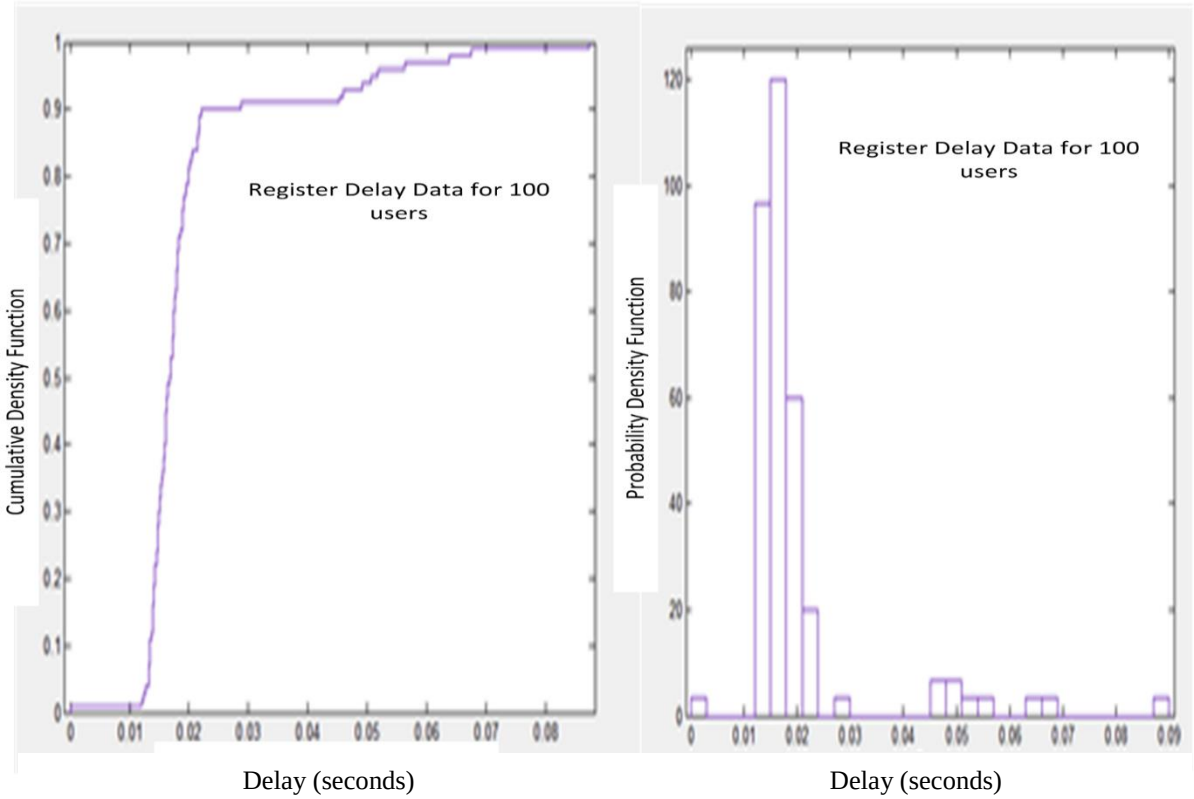


Figure 6.9 CDF and Density functions of the Registration delay for 100 users

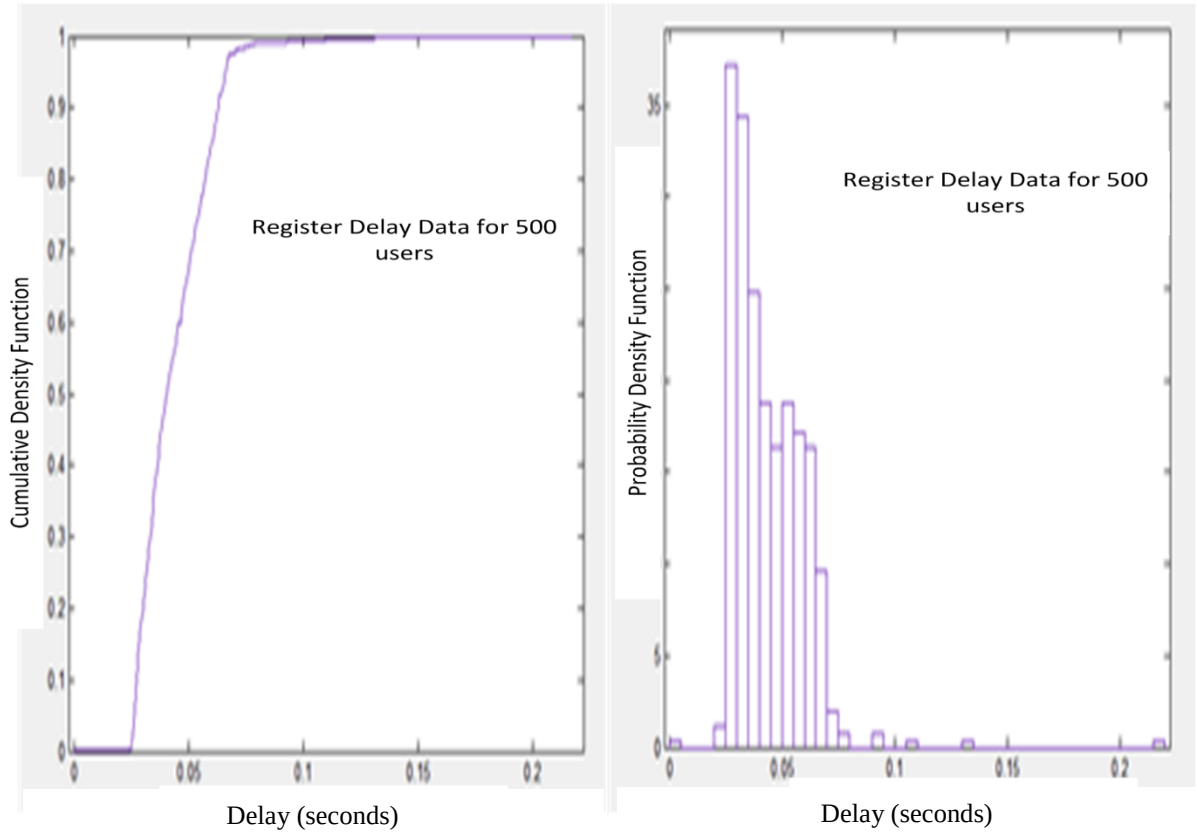


Figure 6.10 CDF and Density functions of the Registration delay for 500 users

From figures 6.9 and 6.10, it is clear that the registration delay for 500 users more than doubled compared to that achieved for 100 users, and increased even further when increasing the number of users in each step. Similarly, seven varying scenarios were implemented using the testbed. In the first scenario, 100 user registration requests were transmitted, while in the subsequent scenarios the number of users was incremented in steps of 200 until 1,300 users were considered in the seventh scenario. To test the scalability of the system, the number of users was gradually increased in order to gain a better understanding of the relation between the number of users and the KPI values for both SIP and IMS.

For more clear threshold values, figure 6.11 shows a smoothed curve of the Probability Distribution Function (PDF) and figure 6.12 the Cumulative Distribution Function (CDF) for the RRD of the first scenario with only 100 users each sending one registration request at a time in sequential order. As shown in figure 6.12, 90% of Registration requests need less than 40 ms to be completed which meets the requirements of mission critical applications and real-time services. As shown in figure 6.11, the highest frequency of the registration trials needs on average 20 ms to be completed. This is considered the best-case scenario and was used as a benchmark for the other scenarios in order to enable comparison of both the RRD time and the percentage of trials that finish at certain time threshold.

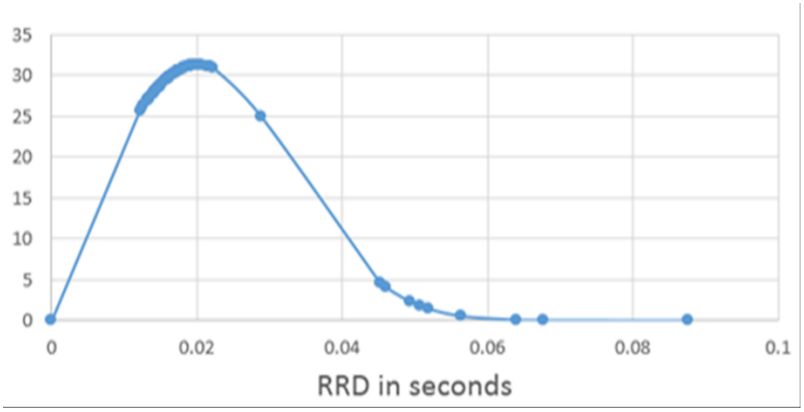


Figure 6.11 PDF of RRD for 100 users

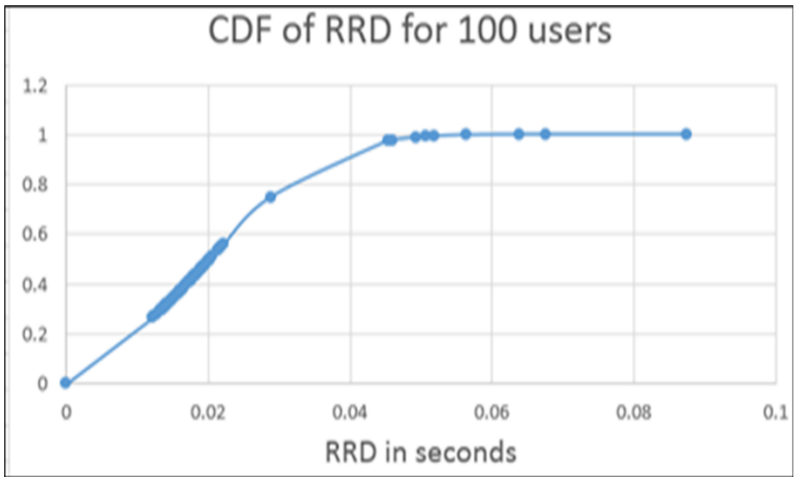


Figure 6.12 CDF of RRD for 100 users.

Similarly, the PDFs and CDFs for all seven scenarios were generated as shown in figure 6.13 and 9. It is clear that when the number of clients increases, the system needs more time to serve the registration requests. This happens due to accumulation of both SIP and DIAMETER signalling messages in the queues of the CSCFs interfaces (especially in S-CSCF) and the HSS interface. Both S-CSCF and HSS are considered bottleneck points of congestion that are affected significantly as the number of registration requests increase. This leads to a queuing delay that emerges rapidly in the system interfaces, which can eventually cause system failure.

Two performance metrics were used to facilitate comparison of the seven scenarios. The first was the time needed to process successfully 90 percent of requests, referred to as 90% completion time (90CT). While the second was the percentage of successfully completed registration requests within 40 ms seconds (which is the maximum RRD time needed to process 90% of requests in the 100-users scenario), referred to as 40 ms Completion Ratio (40msCR).

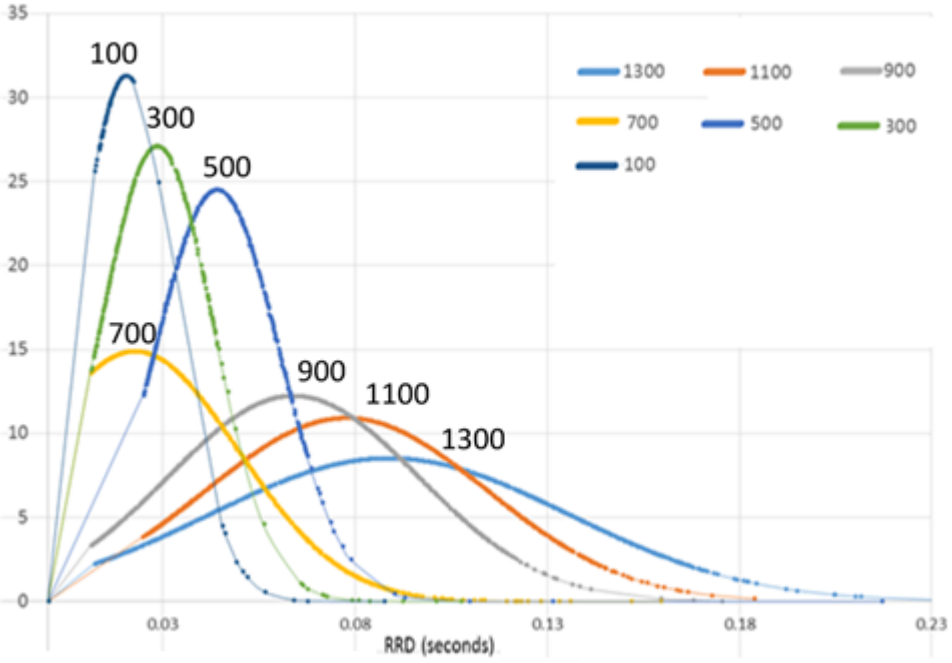


Figure 6.13 PDF of RRD values for all scenarios

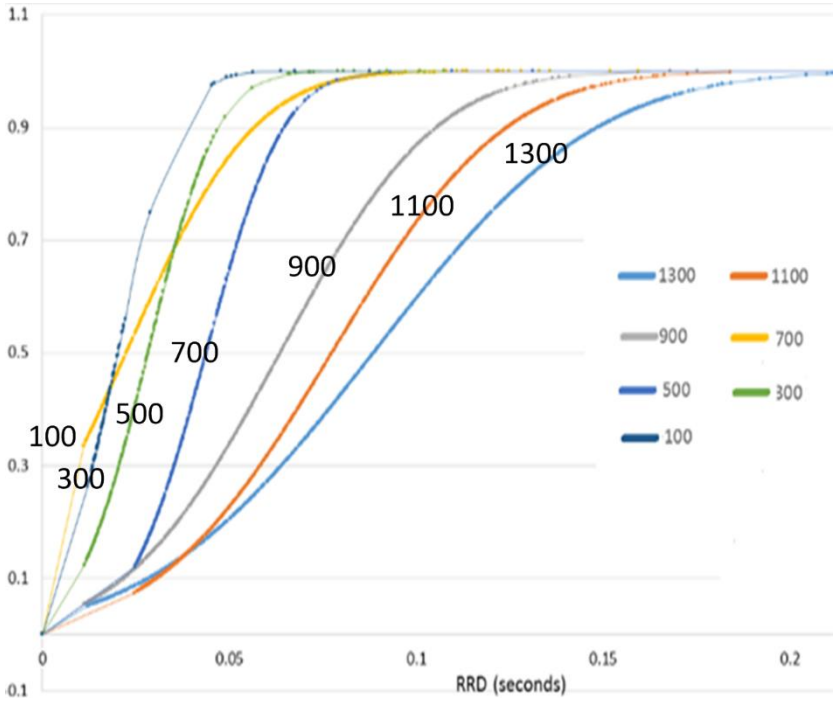


Figure 6.14 CDF of RRD values for all scenarios.

Table 6.3 shows the average RRD values along with 90CT and 40msCR for each scenario and the registration requests per second (RRps) processing rate (which is the number of registration

requests that are successfully processed by the testbed per unit time). The RRps values replicate a real world disaster scenario, where thousands of users may send registration request to gain access to the system - which is supposed to be scalable and reliable - at the same time.

Table 6.3 Calls Statistics from Simulation Results

Scenario no.	RRps	RRD Avg. Value	90CT (ms)	40msCR (%)
Scenario 1 (100 users)	1,800	20	40	100%
Scenario 2 (300 users)	3,600	28	47	80%
Scenario 3 (500 users)	5,900	44	65	40%
Scenario 4 (700 users)	4,900	22	55	73%
Scenario 5 (900 users)	5,500	64	105	23%
Scenario 6 (1100 users)	7,400	77	125	17%
Scenario 7 (1300 users)	5,800	88	150	15%

Based on the results, it can be seen that the RRps increases exponentially as the number of users increases up to a limit (of 1,100 users) before beginning to decrease, leading eventually to system degradation and failure. This is shown clearly by RRps for both scenarios 6 and 7, where the RRps of scenario 7 is much less than the RRps for scenario 6 although the number of users has increased by 200.

As expected, the 90CT increases as the number of users rises, starting from 40 ms for scenario 1 through to 150 ms for scenario 7. This is to be expected due to the increased processing time needed for the additional received registration request. Moreover, it was found that the 40msCR decreases with an increased number of users. Comparing the values with scenario 1 (the benchmark) shows that only 15% of registration requests needed less than 40 ms RRD value to be completed, which again implies that the system is not able to process the received request within very strict time limit.

It was found that, within limits, the system's ability to process the registration requests per time unit increases exponentially when the number of users is increased. However, once the limit is reached, the number of processed requests starts to decrease and eventually degrade leading to system failure. The simulation results show that the system was able to handle a maximum of 7,400 registrations per second, a workload that could occur during a nationwide disaster with many users trying to access the Mission Critical System (MCS).

6.4 PROPOSED FRAMEWORK SCENARIOS SIGNALLING

Following the baseline results presented in the previous section, this section will present the enhanced IMS system evaluation results according to the Framework proposed and suggested implementation presented in the previous chapter. This section will present the sequence diagram of the experiment based on the enhanced framework, then it will show the experimental results based on different scenario assumptions, then it will benchmark the results and compare the performance between different scenarios.

Signalling of control messages is essential part to run the automated experiment. The framework is design to be automated and to run the system without human intervention. To achieve this, there should be a clear definition of the signal timing and sequence order. Sequence diagrams help to better understand the operation and the timing order of the overall system and clarify the scenario assumptions and structure made. The scenarios are summarized in Table 6.4.

Table 6.4 Scenarios Description

Type	Sub-Type	Scenario Number	Number of IMS Cores	Description
Without Using the Framework	**Feed-Forward only**	Scenario (A)	1	System sends requests to single IMS system without using the Intermediate System and its feedback input.
Using the Framework		Scenario (B)	1	System interacts with SUT and Intermediate System (single IMS) but without feedback input from the Intermediate System.
		Scenario (C)	2	System interacts with SUT and Intermediate System (Two IMSs) but without feedback input from the Intermediate System.
		Scenario (D)	3	System interacts with SUT and Intermediate System (Three IMSs) but without feedback input from the Intermediate System.
		Scenario (E)	4	System interacts with SUT and Intermediate System (Four IMSs) but without feedback input from the Intermediate System.
	With Feed-Back input	Scenario (F)	System Defined	System interacts with SUT and Intermediate System. Number of IMS cores is based on the feedback input from the Intermediate System.

6.4.1 Enhanced Traffic Generator:

Before going into the details of each scenario, the Traffic Generator need to be explained in order to fully understand it is capabilities and how it is reflected for each scenario in the following subsections. In contrary to the very basic traffic generator GUI presented in the previous sections, the Enhanced Traffic Generator has more options that is able to represent realistic workloads generated by small set or large set of users. As shown in figure 6.15 there are many parameters and configurations that can be set.

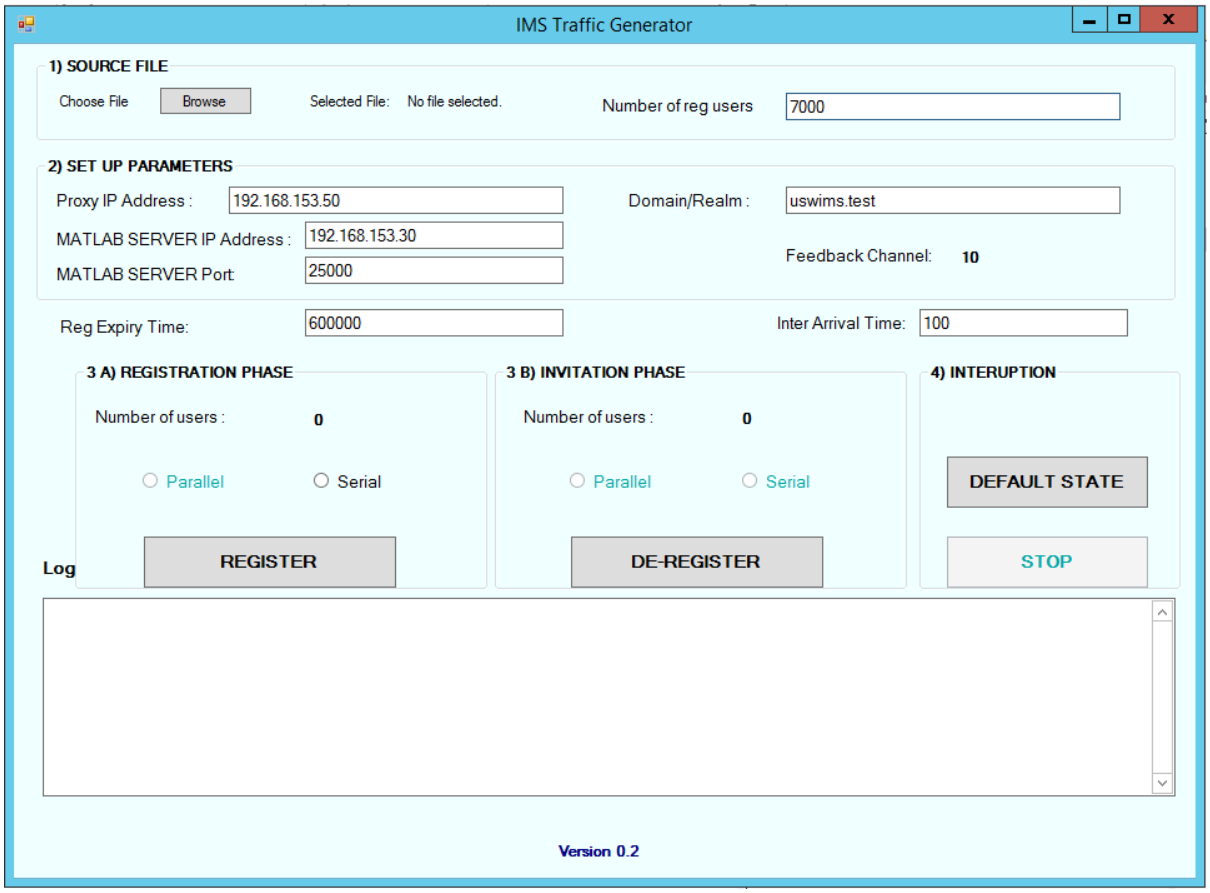


Figure 6.15 Advance Traffic Generator GUI

The test Admin user can select the number of users who will be connected to the core system either manually or by selecting a predefined pre-saved user's profile which is saved in a file in the local disk of the testing machine. Then the user can select the Proxy address (System Under Test Machine IP address) and the MATLAB Server IP Address (the IP address of the Intermediate System), the port at which the UDP data can be sent to the intermediate system can be selected as well. The domain name entered should match the defined domain name of the SUT to avoid redirecting the packets to another domain.

The test admin user can also select the message type (Register or Deregister) and the Timeout value if Register message type was selected. Finally, the inter-arrival time of generated request can be selected as well by the user. The inter-arrival time (τ) reflect the number of generated requested per unit time (λ) in which $\lambda=1/\tau$.

6.4.2 Scenario (A):

In this scenario, the Testing system is connected to the SUT directly without the intervention of the SUT. This Scenario can be used as a baseline to get the best performance estimate of the overall system without using the proposed framework. The signalling Diagram is shown in Figure 6.16 in which it shows the traffic Generator sending predefined number of requests to single IMS system and recording the system response delay locally after getting the reply for each request. For this scenario an internal DNS system is used to minimize the overhead introduced by DNS signalling and to focus more over the IMS signalling performance estimate. Once the timeout is reached, indicating the end of countdown timer, the Generator stops sending requests to the SUT. The recorded delay is an estimate of the SUT performance by calculating the time difference between starting the IMS handshaking signalling process and getting the final reply indicating that the process was finished successfully. Propagation and post processing delay are ignored in scenario and all other scenarios.

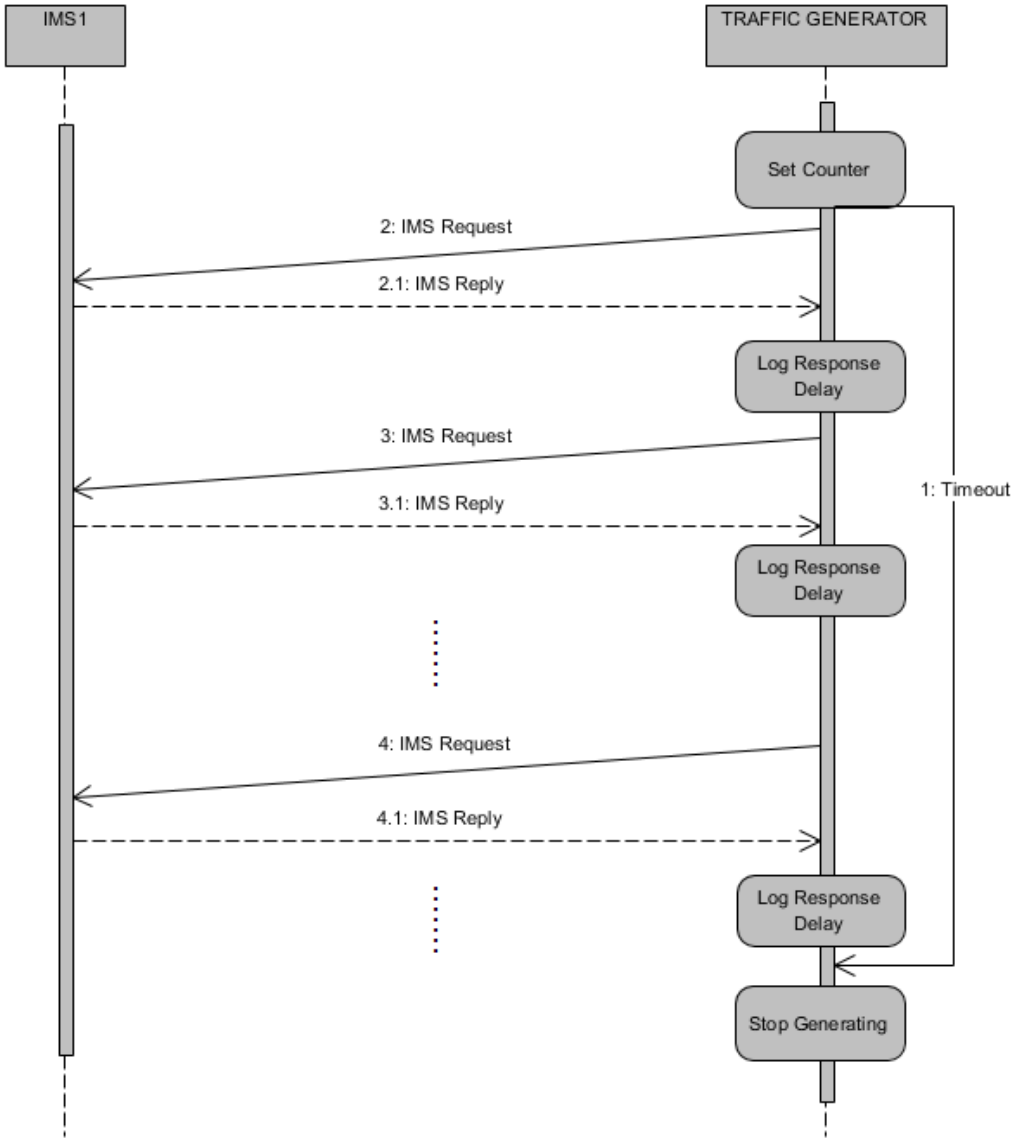


Figure 6.16 Signalling Diagram for Scenario A

6.4.3 Scenario (B)

In this scenario the Traffic Generator (Testing System) sends the traffic to both the IMS (System Under Test) and The MATLAB Server (Intermediate System). Although the intermediate system is used in this scenario, there is no feedback signal fed back to the testing system, such scenarios are referred as Open-Loop or Feed-Forward as indicated in table 6.4. The missing part of the functionality was intended to check the overall system capacity by overloading the system with the maximum stress to get the upper level performance metrics estimate. The SUT will be referred as Single core IMS system to distinguish it from other SUTs within the same group of scenarios. Figure 6.17 shows the sequence diagram of the scenario signalling. The Traffic Generator sends as shown in the figure, for each request sent, the reply

is received by the Testing system and the response time is estimated and sent to the intermediate system for further processing.

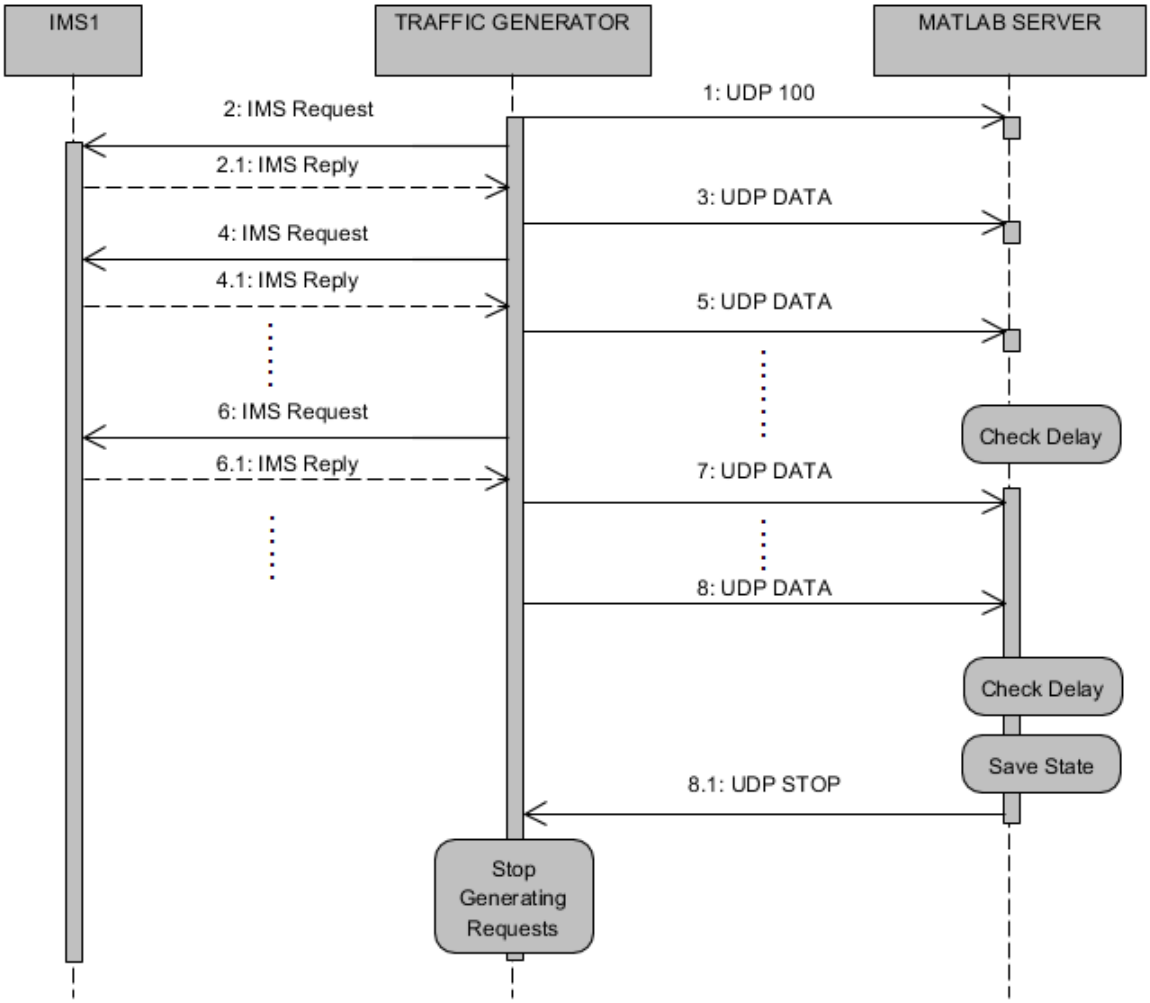


Figure 6.17 Signalling Diagram for Scenario B

The processing part in this scenario is checking if the delay has exceeded the expected values which could only occurs if the message is lost or the IMS server went down. If the recorded delay falls within tolerable limits that reflect the performance of a system running within expected operation mode (even under high traffic volume), then it should be considered as normal delay, otherwise it will send a UDP message to the traffic generator asking for stopping the test after saving the system state and the logged delays.

6.4.4 Scenario (C)

In this Scenario, similar to the previous scenario, the intermediate system is involved but without sending feedback signal to the testing system. However, in this case, the SUT is composed of two IMS systems and is referred as Dual-IMS core system for simplicity. The requests are sent with equal balance to IMS1 and IMS2 cores within the system. and similar to the previous scenario, the responses are recorded and the response time is calculated and sent to the Intermediate system for further processing, the intermediate system evaluate the received UDP data messages loaded with system response delay values occasionally but not continuously to tell whether the maximum system capacity was reached or not. If the delay exceeds delay tolerance for normal operating system (response takes seconds for example), the intermediate system will send a message via UDP to the traffic generator asking for stopping the test. Figure 6.18 shows the signalling diagram described for this scenario.

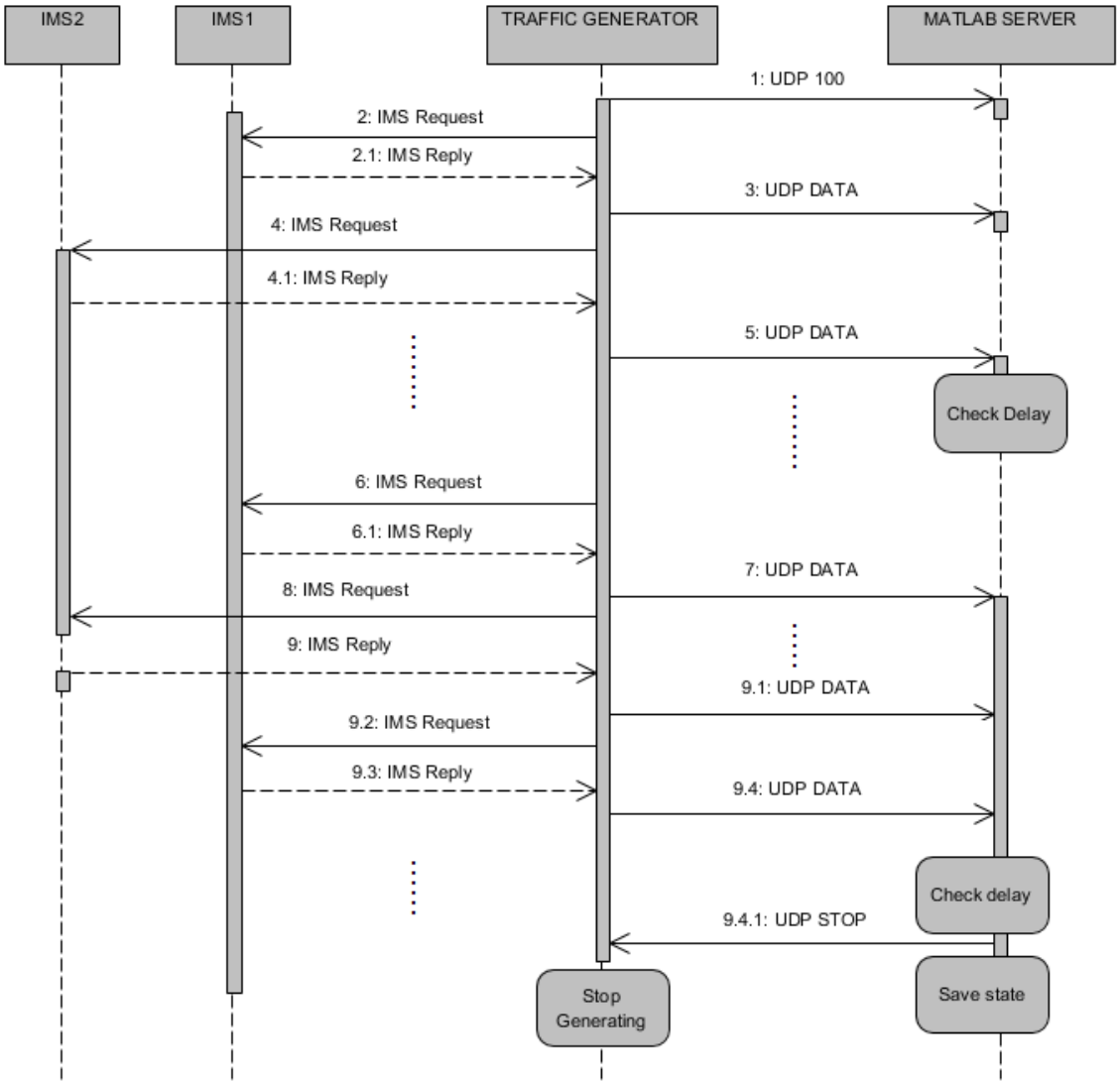


Figure 6.18 Signalling Diagram for Scenario C

6.4.5 Scenario (D)

Similar to the previous two scenarios, there is no feedback signals from the intermediate system although that the intermediate system (represented by the Matlab server) is involved. As shown in figure 6.19, the Testing System sends the requests equally to the three IMS servers and then get the replies and record the delays similar to the way done in the previous scenarios. The delay check is done as well similar to the previous two scenarios.

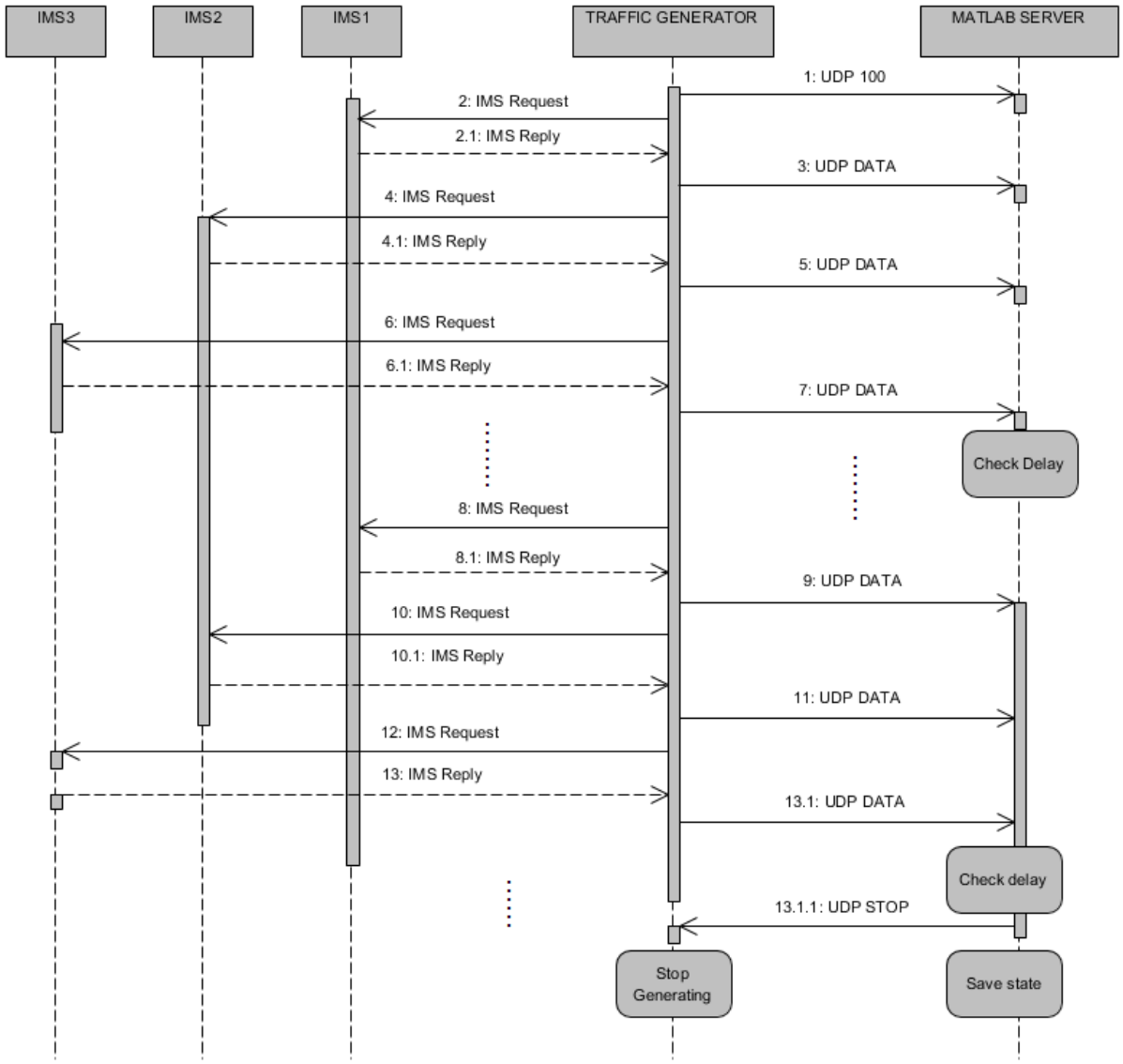


Figure 6.19 Signalling Diagram for Scenario D

6.4.6 Scenario (E)

Similar to the previous Scenarios (Scenario B, Scenario C, and Scenario D) the requests are sent to the SUT but this time it has four IMS systems which will be referred as Quad-Core IMS system. All the responses and delays are recorded in similar way. The load is equally balanced among the four cores as well. Figure 6.20 shows the signalling diagram for this scenario.

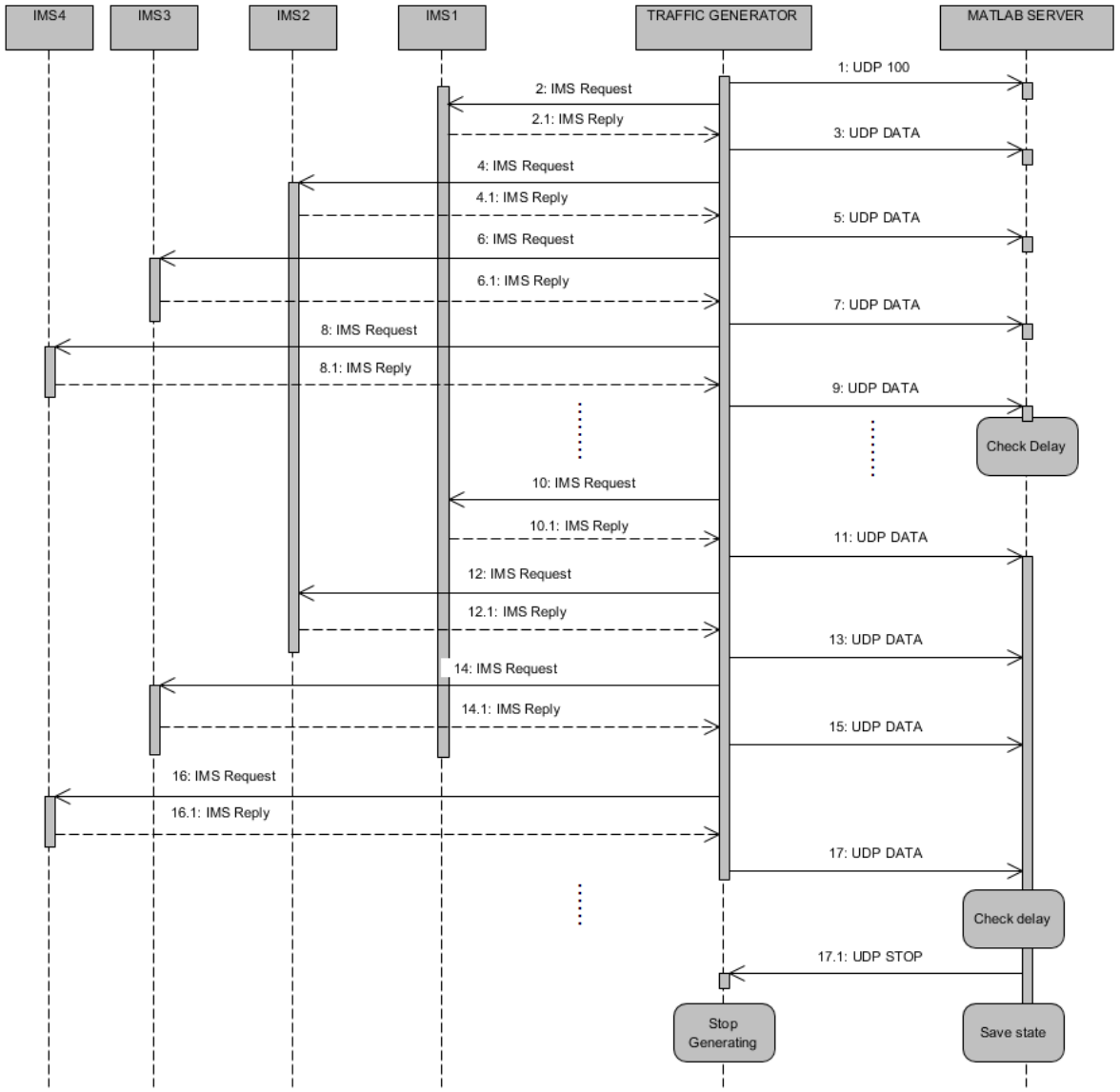


Figure 6.20 Signalling Diagram for Scenario E

6.4.7 Scenario (F)

For this Scenario, all system entities are involved similar to the way it was described in the proposed framework in chapter 5. The feedback channel is fed to the testing system which decide the next SUT setup based on the received feedback channel value as indicated in table 6.5. In contrary to the Feed-Forward (or Open-loop scenarios) mentioned previously, this scenario does not have a fixed setup of signalling due to the dynamically changing selection criteria of the core IMS components, and as mentioned previously, only the "BASIC" set of feedback channel values will be considered for this research. Therefore, it would be easier to start with single IMS core SUT and show the signalling diagram as shown in figure 6.21, then demonstrate other possible continuation of the scenario.

Scenario F.1 which is a special case of this scenario shows the Traffic generator sending a UDP message loaded with control message (100) to the intermediate system indicating that there is data will be received shortly, the intermediate system will start running a packet sniffer to save the pre-processing overhead. Once requests and replies are sent to the SUT (which has only one IMS core in this scenario) the delays will be calculated and sent to the intermediate system similar to the way described in the feed-forward scenarios described before. Upon receiving the first UDP packet data loaded with the response delay, the intermediate system will receive all subsequent responses and save it as one single batch of data for further processing. The data will be analysed and processed based on certain criteria that will be shown in the next section. Based on the criteria output, it will be decided how to proceed, if the criteria condition is not met (i.e. indicating that the threshold was not reached yet) the intermediate system will start another sniffing iteration after sending FeedBack Channel (FBC) update message, the FBC value should be similar to the current value ("1000").

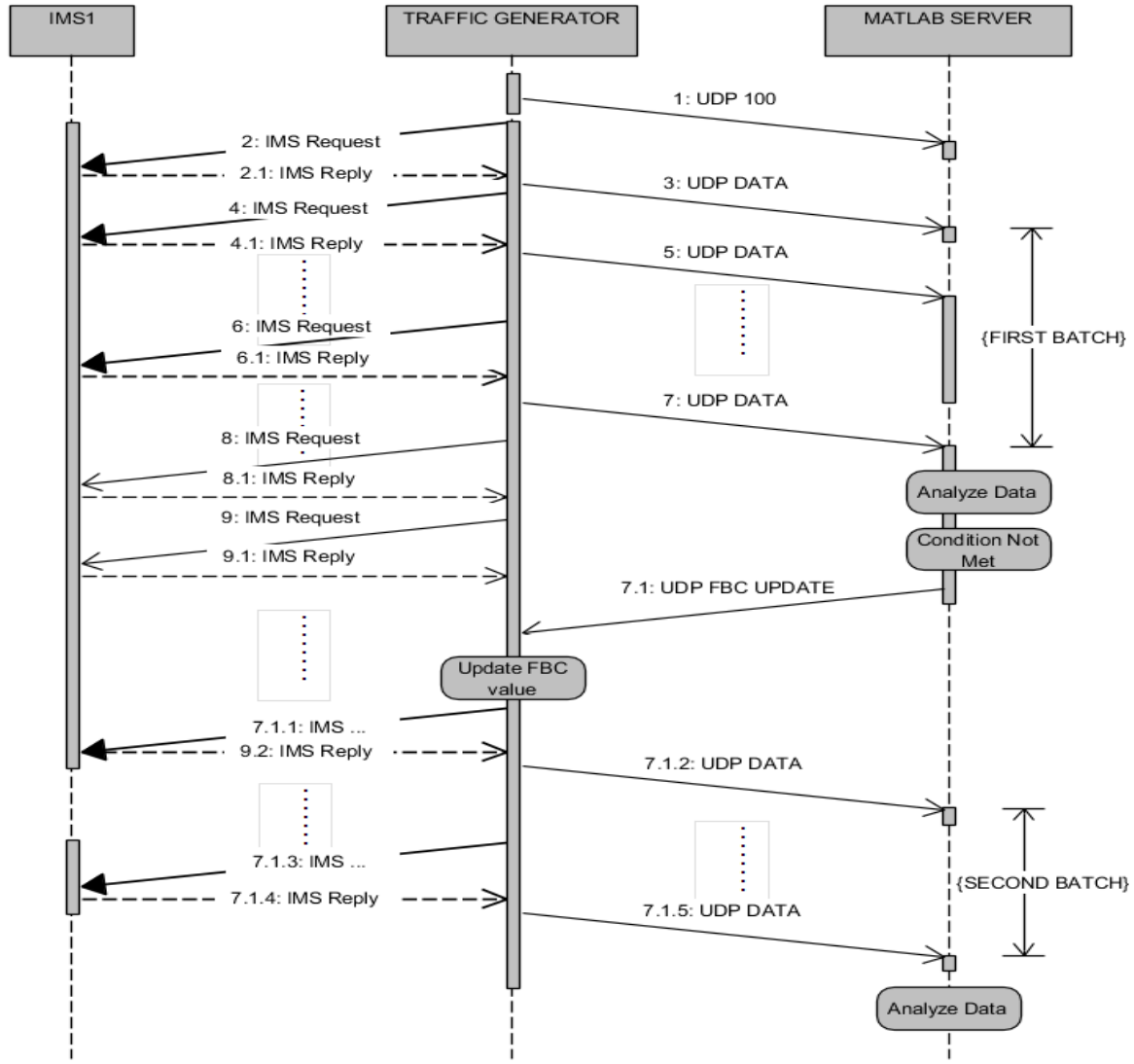


Figure 6.21 Signalling Diagram for (Scenario F1)

The process will keep repeating until getting a batched processed and analysed and the condition is met in the set criteria, then it will send UDP FBC update message to the testing system changing the FBC value from "1000" to "0100" which indicates that the second IMS core need to be involved in the process. As figure 6.22 shows, Upon receiving the UDP FBC update message by the Testing system following a successful meeting of the set condition, a new stage will start (Referred as Scenario F.2) the new requests will be forwarded according to the FBC value to the corresponding IMS core components. The handover between the two scenarios is done smoothly and without noticeable overhead due to the way the testing system keep monitoring the FBC values before sending requests to the SUT.

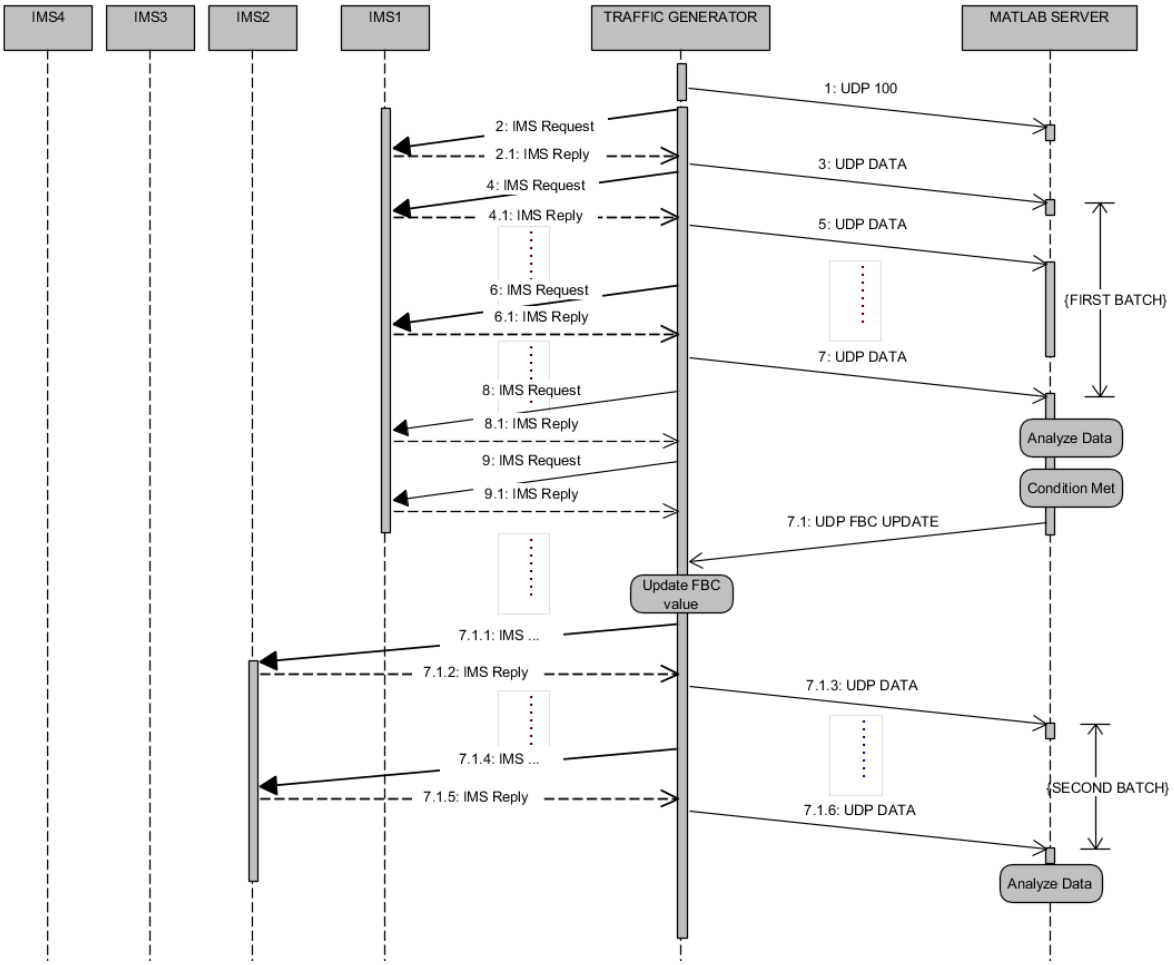


Figure 6.22 Signalling Diagram for (Scenario F2)

Following the successful update of FBC to a new value and sending the requests to the new IMS core component, a new challenge arises here in managing the two Home Subscriber Station (HSS) databases. According to the previously described proposed framework, there is separate HSS database for each IMS core entity, this simplified design architecture comes at the expense of added complexity in data consistency for all databases. Consistency is essential to ensure that the user testing system and the intermediate system does not need to keep track of user's registration state and where the profile record is saved, this realistic assumption need

to be maintained to reflect a real workload traffic processed in the real world core network regardless of the originating source. To overcome this challenge, all HSS databases were connected via Diameter/SQL signalling domain to a single database management entity running a database management tool and ensure updating the database continuously. A dedicated server running a database management tool and monitored via PHP based interface by admin user is used during the experiment setup. The human intervention is for monitoring purposes only to ensure that the databases were updated successfully and has no effect over the automated running nature of the scenarios. Thanks to the separate dedicated signalling domain for database management, it has no or minimal added overhead to the running scenario, not to forget mentioning as well that the consistency enforcement policy does not need to run periodically over short period intervals, depending on the load and database changes, it may run following or before FBC update cycle.

Figure 6.23 shows the database update cycles following successful FBC update and sending the requests to the new IMS core. As shown in the same figure, Scenario F.3 start sending to IMS3 core after getting a new FBC update message (which is "0010" in this case) and then forwards the new arriving requests to IMS3 core followed by HSS Databases update cycle for HSS1, HSS2 and HSS 3. Similarly, for Scenario F.4, following receiving FBC "0001", the Traffic generator sends the new requests to IMS4 and then update all databases (HSS1, HSS2, HSS3, and HSS4) accordingly to ensure integrity.

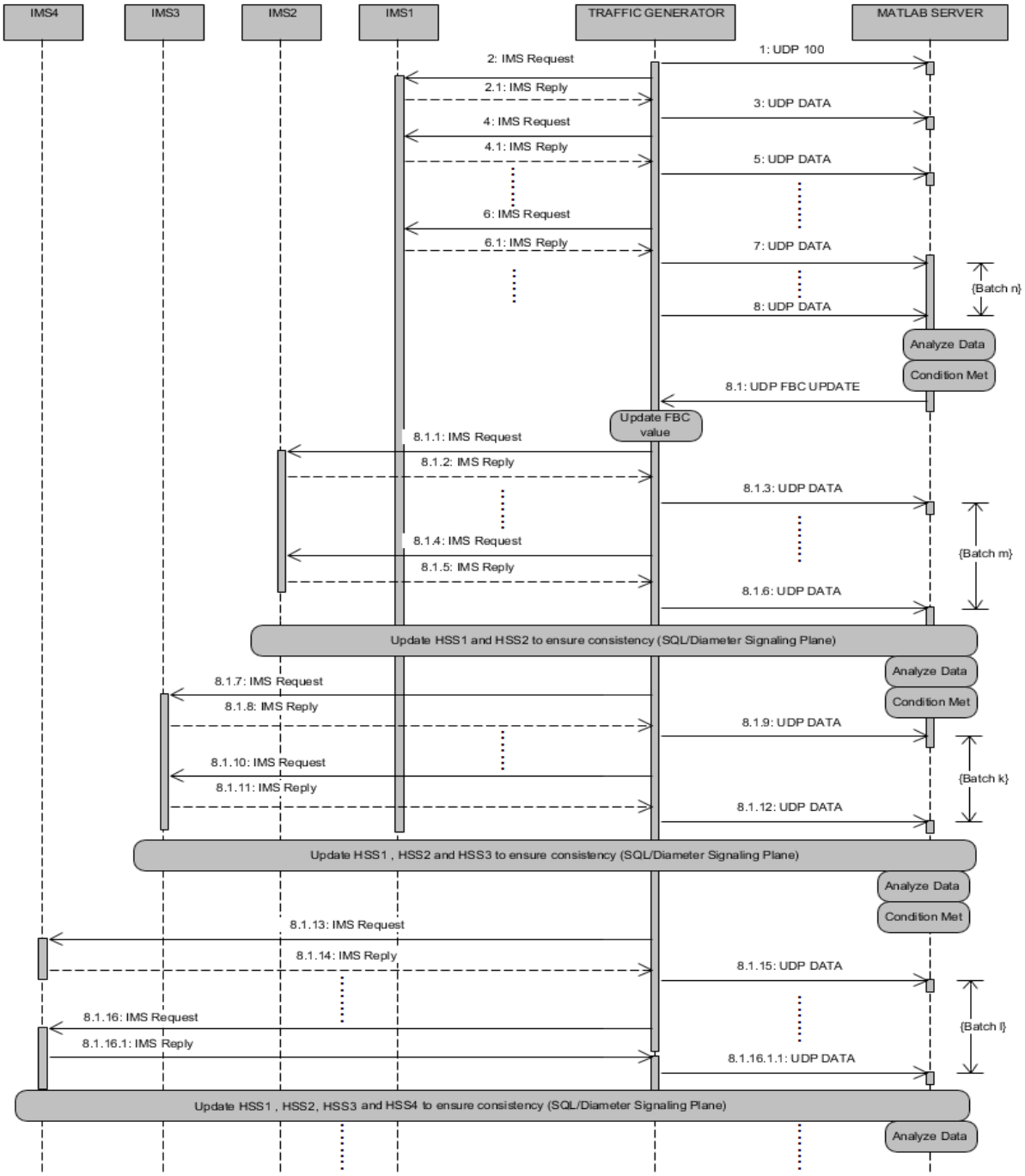


Figure 6.23 Signalling Diagram for (Scenario F3)

6.5 SCENARIOS RESULTS EVALUATION

I this section, the evaluation results of the experimental scenarios presented in the previous section will be demonstrated and explained. As mentioned earlier in the previous section, the main objective of the scenarios is to understand the system performance and capacity limits under different load stress values. Based on Scenario An experiment, in which the testing system was connected directly to the system under test, different scenario runs were carried out to cover most of the scenario options. As figure 6.24 shows, eighteen experiment runs was done while varying the registration request rate to ensure the accuracy of result and to minimize the running environment potential errors. In each run, the SUT was tested against the maximum number of served users (System Capacity) before the system degrades and becomes unresponsive, the duration of the entire test was also measured as an indication of system ability to process the requests in a timely manner.

The end of the test in which the timer is stopped is based on the following assumptions:

1- The SUT returns SIP Server Time-Out message (SIP 504) which indicates that the timeout at the server side has reached without being able to process the response.
2- SUT returns SIP error forwarding message (SIP 500), in which the PCSCF server is unable to forward the request to the next S-CSCF assigned server.
3- SUT returns Busy everywhere message (SIP 600). In which the SUT indicates that the SCSCF server has crashed and unable to process any request.
4- The system enters an idle state without returning any messages, this mainly is due to a breakdown of the PCSCF server and not being able to process the received request.
5- If the request message get congested in one of the networking nodes (switches or routers) and not being delivered into the SUT.

If any of the previous scenarios happen, the test is considered failed and need to run again, or finished successfully. For the 504 SIP, 500 SIP, 600 SIP error messages in addition to becoming unresponsive to further requests, the test is considered successful and measurement are recorded. However for the last case, in which the network becomes congested, the test is considered failed scenario and the entire test is repeated after resetting the whole system components. The same assumptions are considered for the other experiment scenarios that will be presented later.

6.5.1 Scenario (A) Results

Based on the results shown in figure 6.24, increasing the request rate (λ) reduces the end time of the experiment run exponentially. And has relatively no effect over the overall system capacity. Based on the statistics calculated, 235 seconds was the average end time of the experiment run with 2878 average number of served user before the system breakdown.

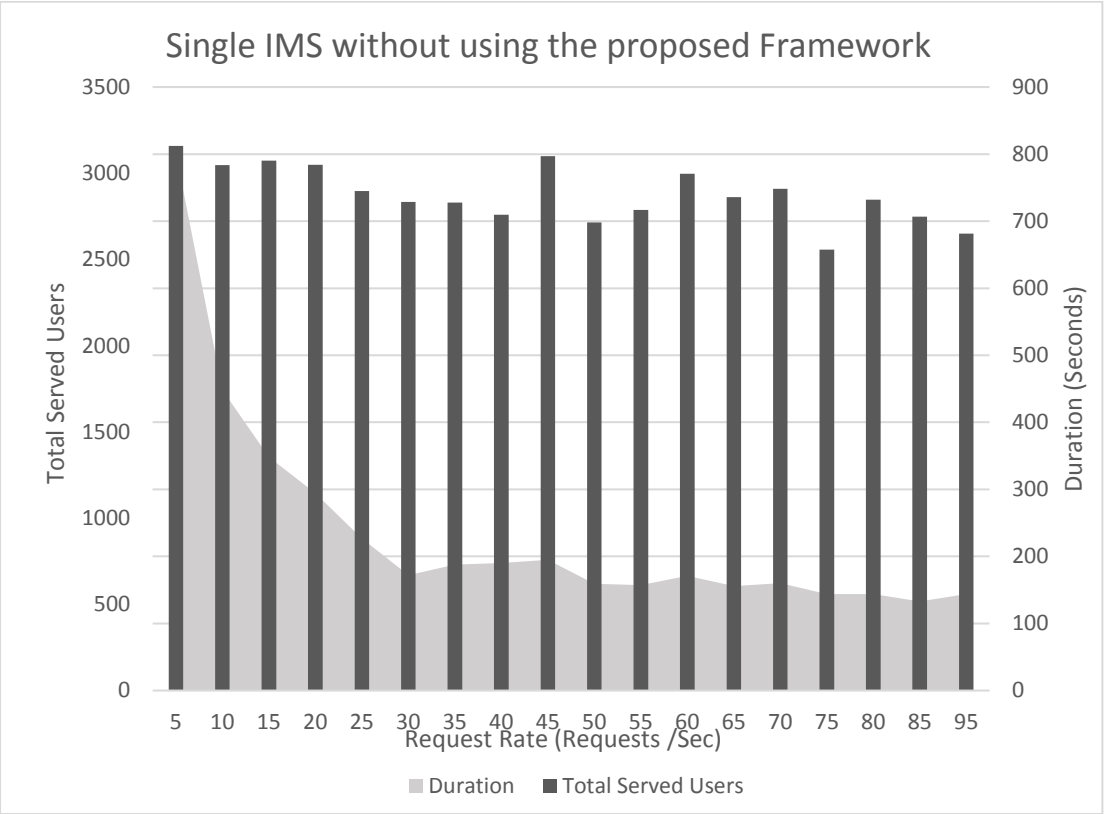


Figure 6.24 Single IMS without using the framework Results

To get a better estimate of the performance estimate for each run at the transaction level, the log files at both the testing system and the SUT were analysed to get processing rate of the request. In contrary to the total time each run takes until it finishes, the processing rate of the requests gives an indication of the system ability to handle the workload before and after entering performance degradation point. Figure 6.25 shows the request processing rate for all the experiment runs. The rate was calculated according to 200 requests window size for each run. The result shows clearly that the system processing rate tend to saturate and then degrades , statistics shows that the saturation point is reached at (Run 14), in which the system processing rate becomes 20 processed requests per second. While increasing the arrival rate in the subsequent runs, the ability of the system processing the new and older arrivals remains the same, which leads eventually to system degradation due to the quick building up of request inside the system as shown in experiment run point (Run 18).

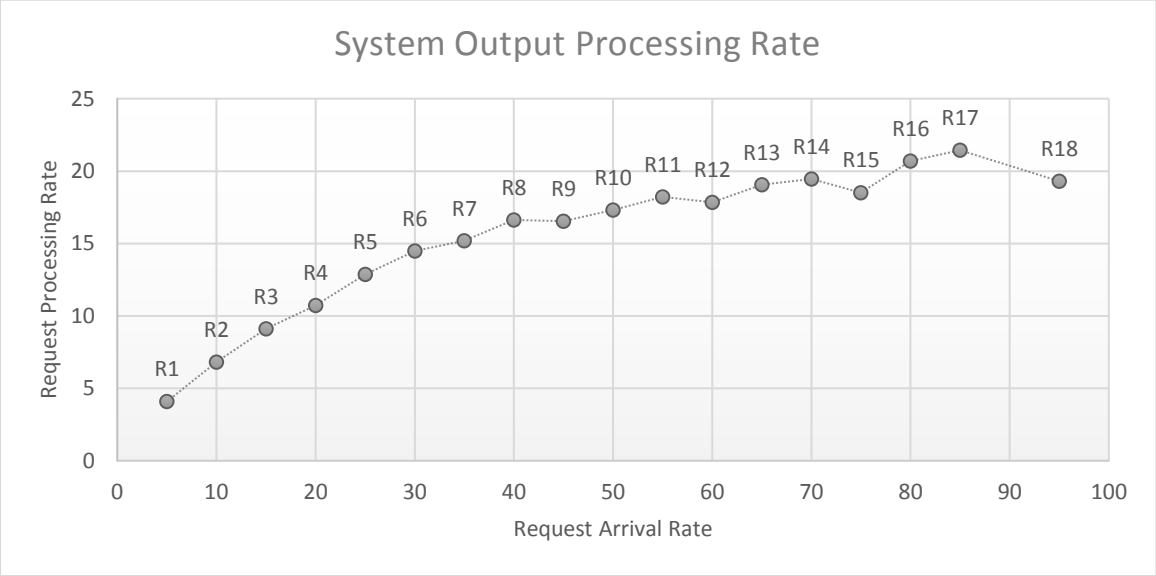


Figure 6.25 System output processing rate

Another comparison between the arrival and processing is shown in figure 6.26. The increasing gap between the two rates is increasing while increasing the arrival rate for each run. The difference between the two rates contributes in filling the queues inside the system and cause system breakdown.

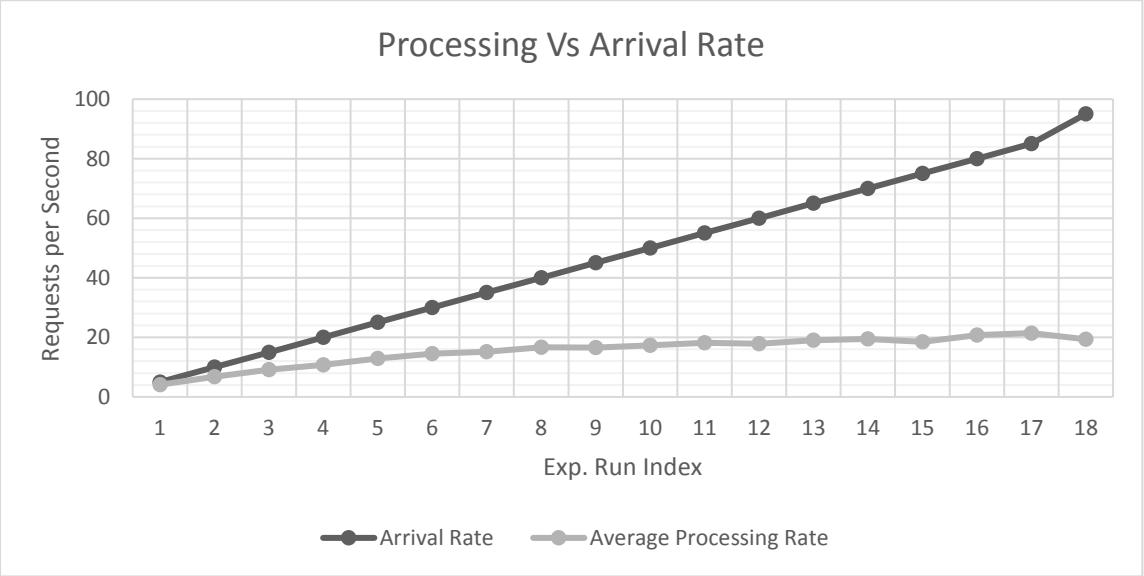


Figure 6.26 processing vs arrival rate

6.5.2 Scenario (B) results

Figure 6.27 shows Scenario B experimental results. Similar to Scenario A results, the Capacity and total test time was measured for each experiment run. While increasing the request arrival rate for each run, the total time of the test for each run before reaching the breakdown point decreases and the total capacity remains relatively unchanged. From one side, the average running time was 1162 seconds which is much more than the previous scenario, and fro the other side, the average total served users was 3110 which is a slight enhancement is overall system capacity compared with the previous scenario. The main difference between Scenario A and Scenario B is applying the proposed framework as an open-loop system without enforcing the feedback policy. So, the added delay is mainly due to the signalling overhead introduced by the intermediate system.

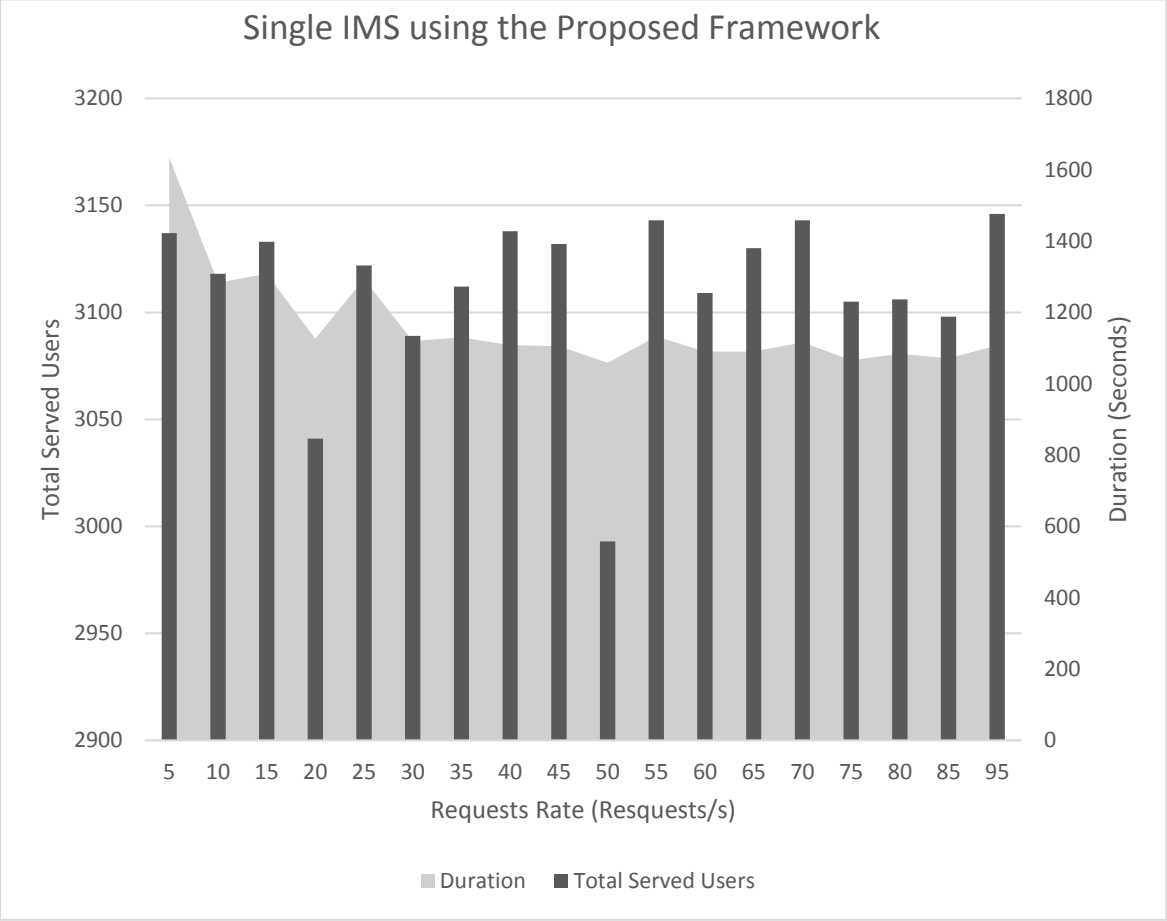


Figure 6.27 (Scenario B) Results

Figure 6.28 shows average requests processing rate for each run, and figure 6.29 shows the difference between the arrival and processing rates. Ignoring the first run value (R1) in which the DNS signalling take place only at the beginning of this run, the remaining values are

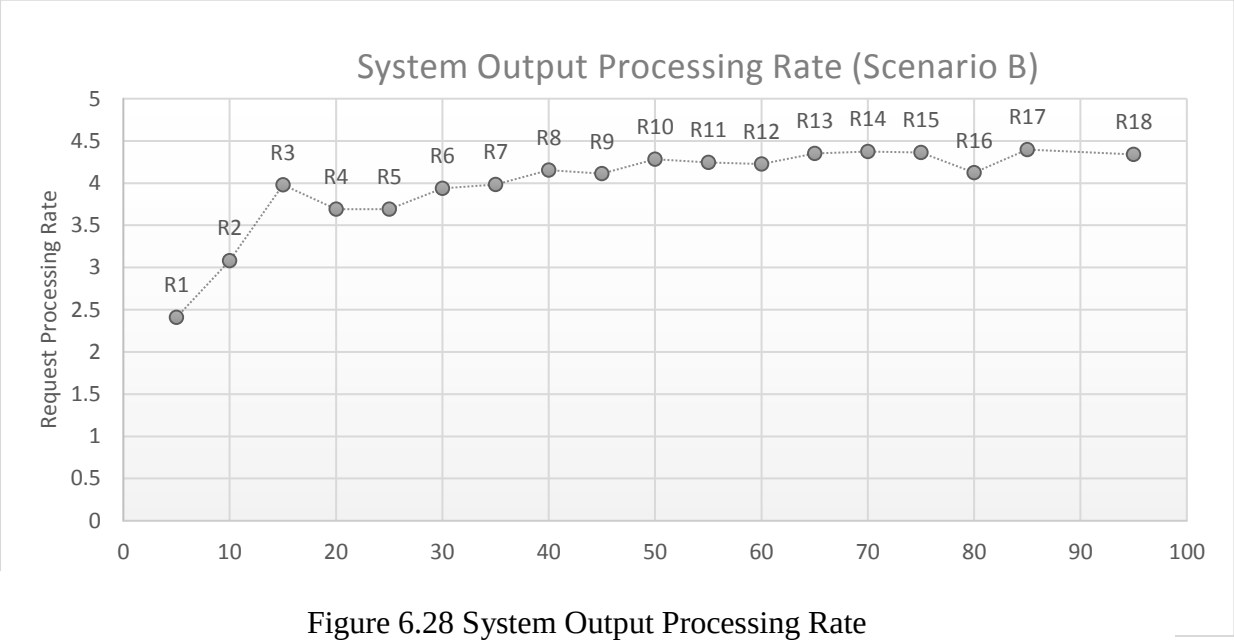


Figure 6.28 System Output Processing Rate

relatively close to each other regardless of the change in the arrival rate. This becomes more obvious in figure 6.29 that shows the difference between arrival and processing rates.

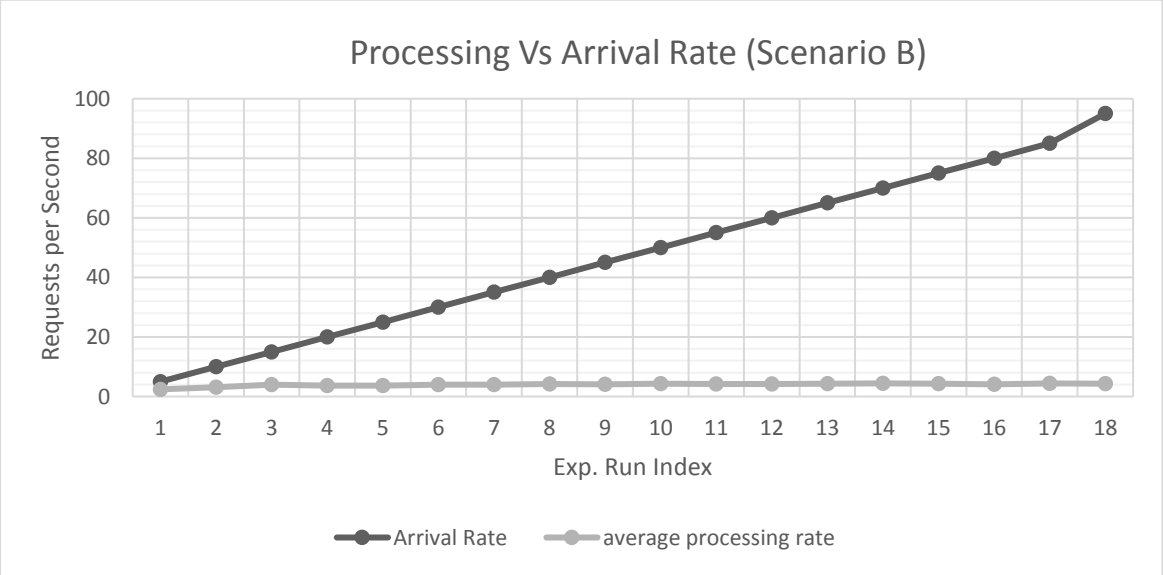


Figure 6.29 Processing Vs Arrival Rate for (Scenario B)

6.5.3 Scenario (C) results

In this scenario, the framework is used in open loop mode and the load is distributed among two core IMS system. Figure 6.30 shows the multiple scenario runs, the capacity and total experiment time was calculated for each run.

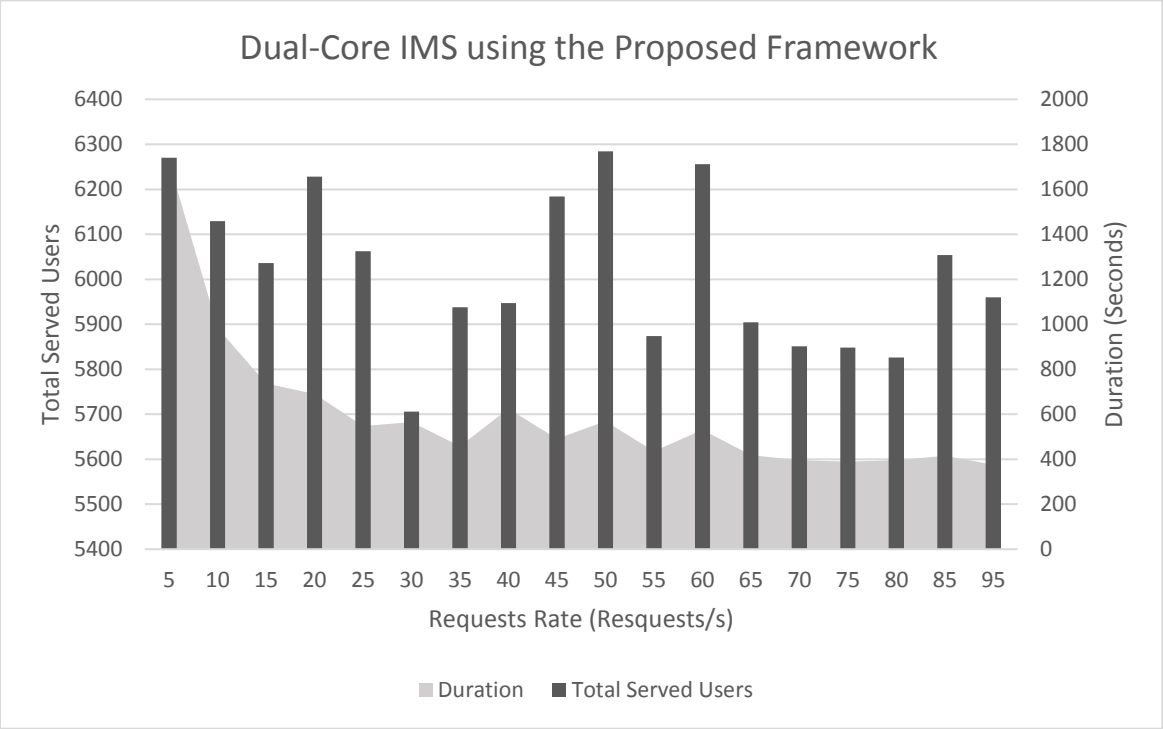


Figure 6.30 (Scenario C) Results

The total served users (system Capacity) was 6020 in average, and the average duration of system run time before reaching the breakdown point was 596 seconds in average. More about results discussion will be explained in the next sections. The system requests processing rate is shown in figure 6.31 and the difference between arrival and processing rates is shown in figure 6.32.

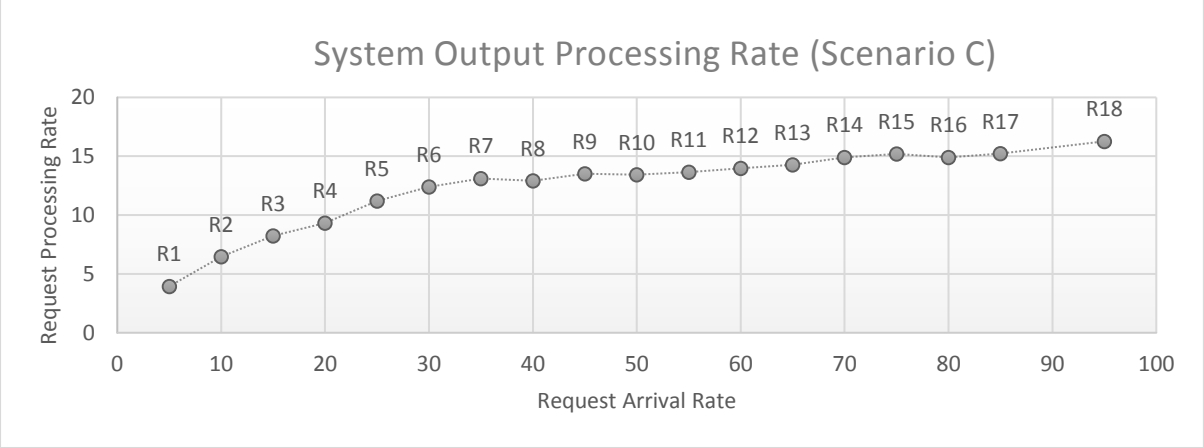


Figure 6.31 System Output Processing Rate

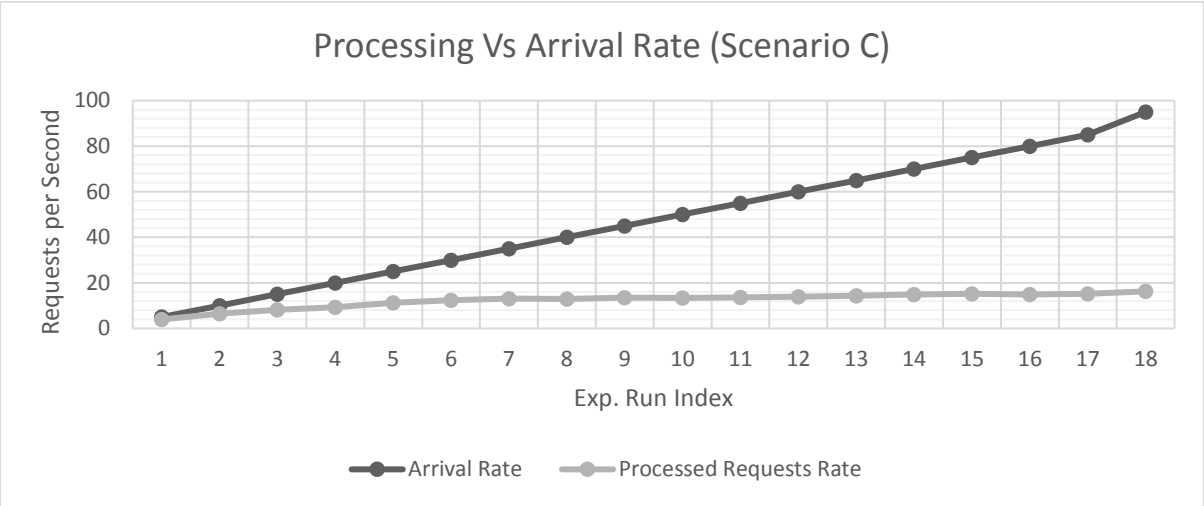


Figure 6.32 Processing Vs Arrival rate

6.5.4 Scenario (D) Results

In this scenario, three core IMS system was deployed using the proposed framework as an open-loop system. Figure 6.33 shows the capacity and total delay measurements. As expected, the total system capacity was extended with an average of 9089 served users for all runs, the total serving time for each run was 877 seconds. Figure 6.34 shows the processing rate of user requests, and figure 6.35 shows the deviation between the arrival and processing rates, the values were collected within a window size of 400 requests due to the large size of log files.

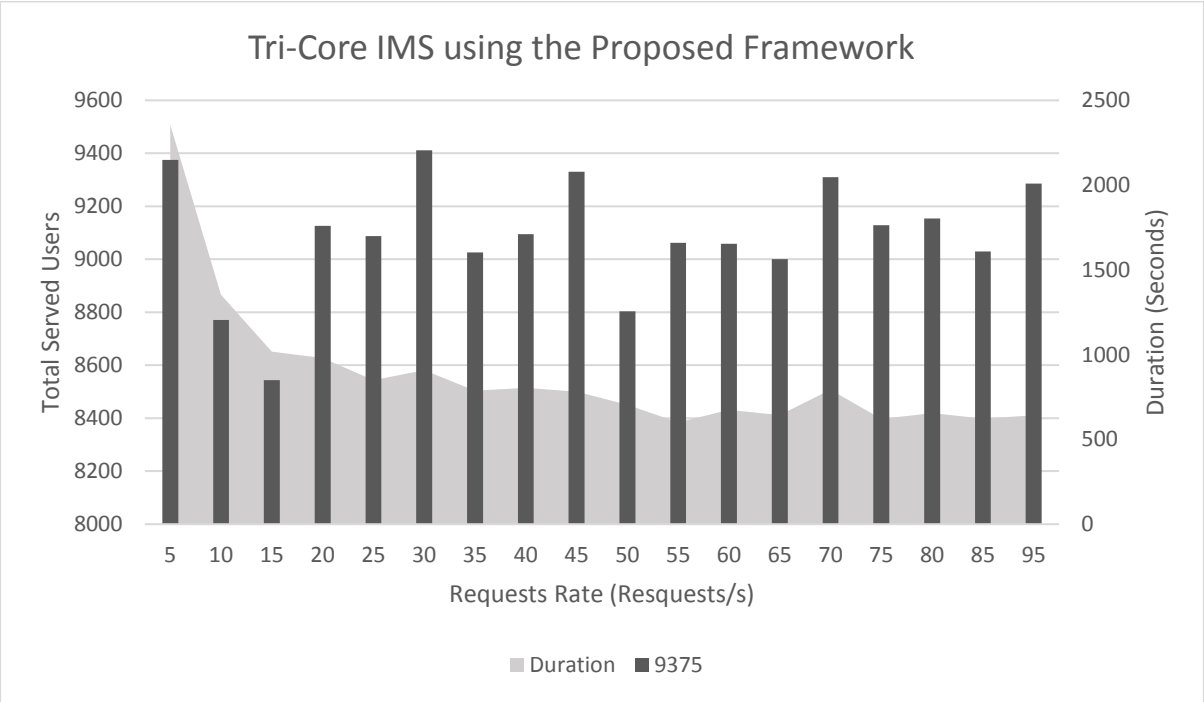


Figure 6.33 (Scenario D) Results

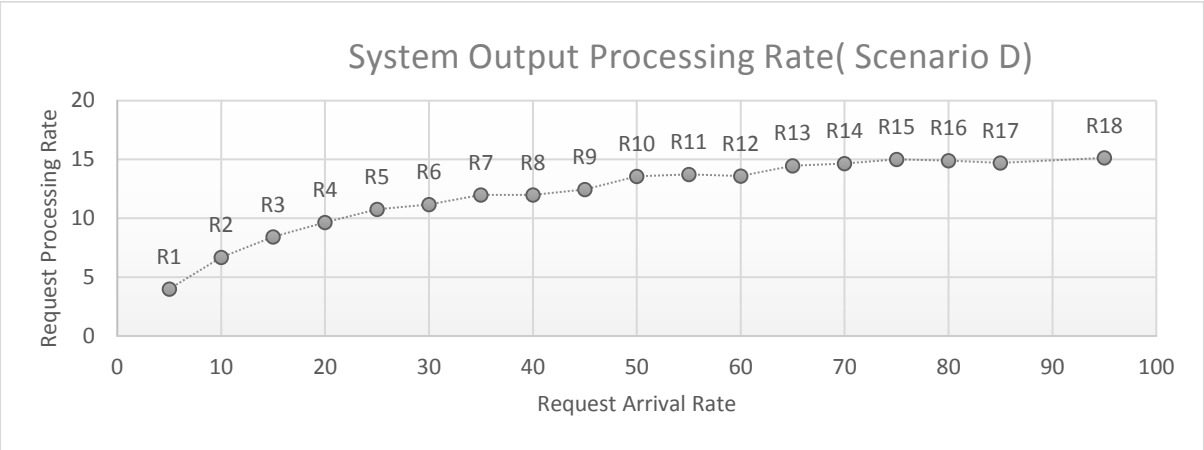


Figure 6.34 System Output Processing Rate

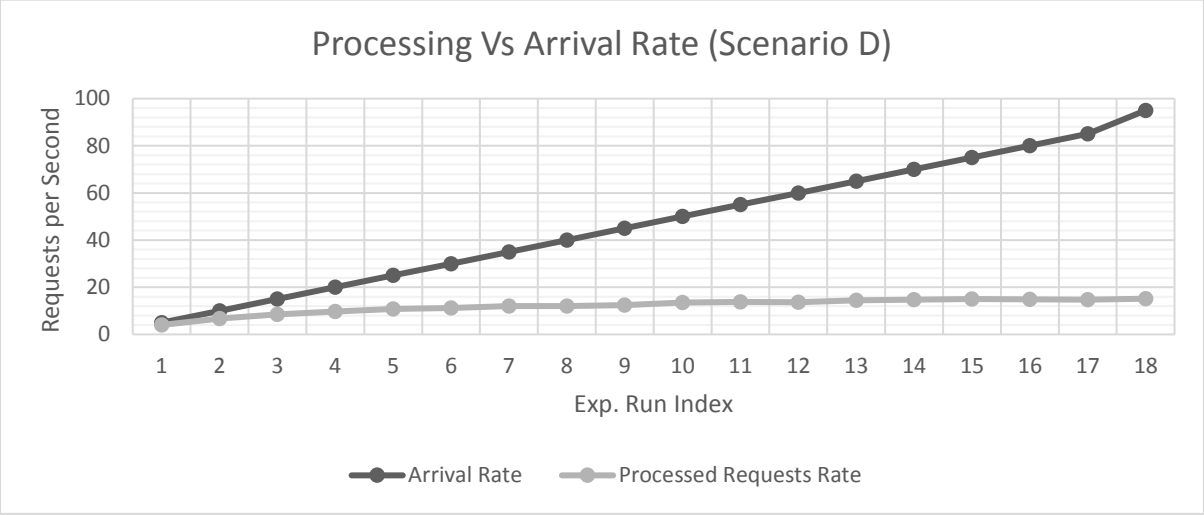


Figure 6.35 Processing Vs Arrival Rate

6.5.5 Scenario (E)

This scenario was not implemented for the following reasons:

1) The Scenario needed more computational resources that exceeded the available ones.
2) The limit of the registered users in the system database was set to 10000 users, Scenario D almost reached 90% of this limit. Therefore, this scenario is expected to exceed the limit of the number of registered users and a new setup then will be needed.
3) The statistics and results collected from the previous scenarios was enough to get an estimate of the system performance. Increasing the number of IMS cores may not add significant amount of information.

For the previously mentioned reasons, it was decided not to implement this scenario. However, it will be considered as part of the future work in which more analysis on the multi-server system will be carried out.

6.6 RESULTS DISCUSSION AND ANALYSIS

Following the results presented in the previous section, a better comparison and benchmarking can be made of the different scenarios. As figure 6.36 shows, the average experiment running time (in seconds) before reaching the breakdown point for the four scenarios. Scenario A (without having the intermediate system) was the least, and the remaining scenarios time was decreasing with the increasing number of IMS core components. The total delay gives an indication of system reliability, availability and ability to tolerate high traffic loads, but it needs more metrics to get an accurate performance estimate of the four scenarios. Therefore, having a system connected directly to the testing system (Scenario A) is considered as a bottom line system with the least availability and scalability potentials compared to the other systems.

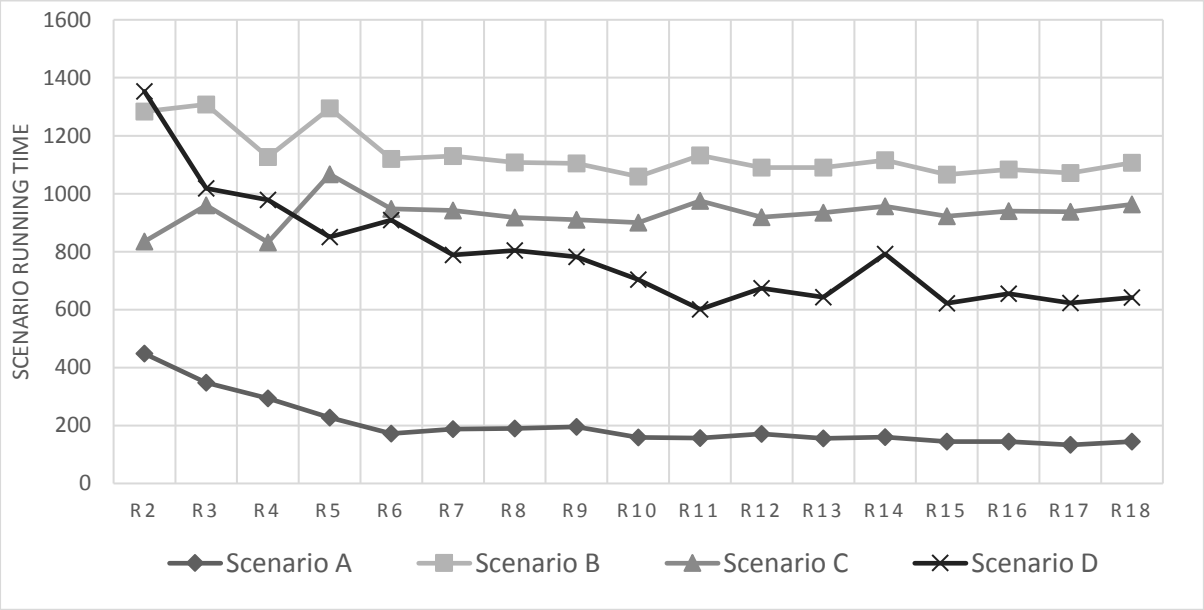


Figure 6.36 Average experiment running time for all Scenarios

The previous figure, showed the variations of total system availability while varying the arrival rate of requests. Figure 6.37 shows the average overall system availability time for each scenario, it is also compared against the overall successfully served users as an indication of system capacity. As mentioned previously, the overall system availability was the least for scenario A, but as figure 6.37 shows, the system capacity was the least as well for the same scenario. In the other scenarios, the system capacity was increasing linearly with the number of subsystem IMS cores deployed, the maximum was for scenario D in which three IMS cores was implemented. Interestingly, the overall system availability was the maximum for Scenario B with less values for the other two scenarios. A new performance measure was needed to link between system efficiency in processing the requests and system availability, having an all-time available system with low efficient processing is not considered the best option.

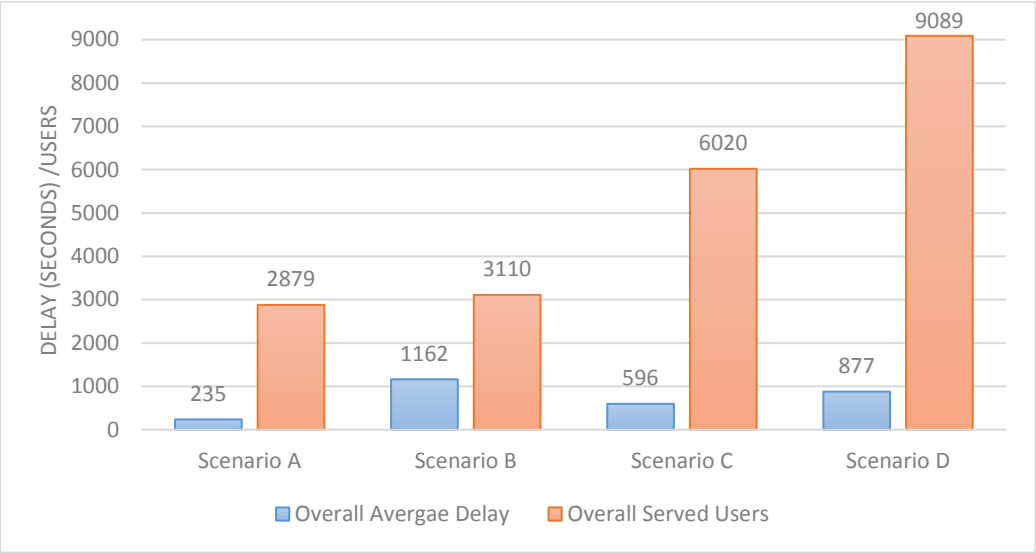


Figure 6.37 Average overall system availability time for all Scenarios

Figure 6.38 shows the requests processing rate for all scenarios, having a testing system connected to the IMS subsystem (Scenario A) had the best performance due to the elimination of the added signalling overhead described in the sequence diagram in the previous sections. The enhanced processing rate comes at the expense of short system availability time. For the other scenarios, the same figure shows that both scenarios C and D has nearly similar processing rate with slight advantage for Scenario D, this is expected due to the difference between system running time of both scenario. Refereeing to Figure 6.37 shows that although Scenario D has higher capacity compared to Scenario C, but Scenario C has a less Running time that makes the ratio between total users and average running time for both scenarios almost identical. Taking into account that the aforementioned ratio is a representation of the average overall processing delay, the processing delay for both scenarios will therefore the same. And finally the least processing rate was for Scenario B which is considered the worst case scenario in this case.

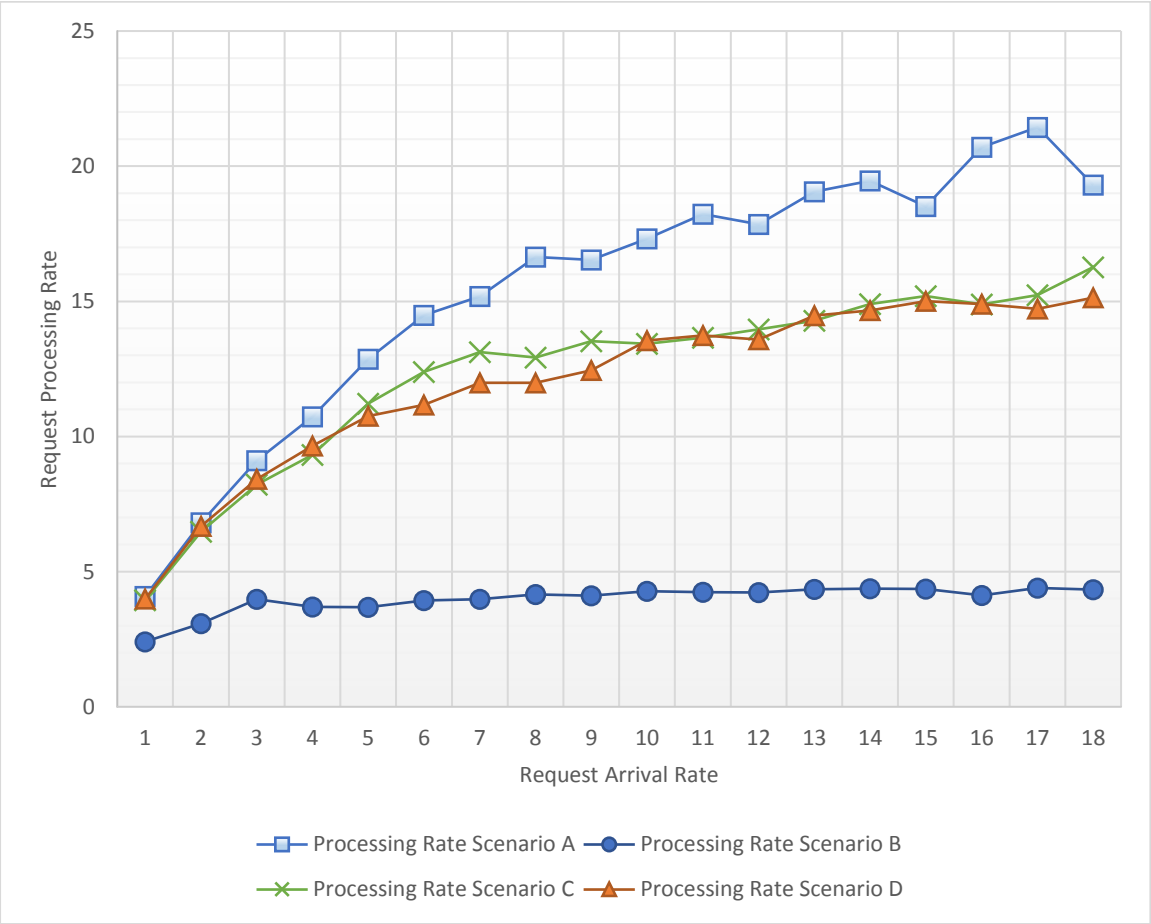


Figure 6.38 Requests processing rate for all scenarios

Comparing the processing delay against the arrival rate (which is the same for all scenarios), shows that there is a considerable difference between the processing rate and the arrival rate for all scenarios. As mentioned previously, this difference contribute in building up the accumulated requests in the queues of the intermediate system and cause a system breakdown at the end. The difference between processing rate and arrival rate is denoted as Δ and shown for each scenario in figure 6.39. Δ2 and Δ3 are almost identical.

As described in the framework design, system will analyse both the forward requests traffic and backward response traffic to get the rate for each direction. Both values will be used to decide the threshold that distinguish one scenario from the other. However, in this case, it is difficult to have a distinction between scenario C and Scenario D performance due to nearly same processing delay, this creates complexity in the system operation if a clear definition of the current system state and the next intended operation state need to be clarified. It worth mentioning at this stage that the system state is fully defined by the number of core IMS active entities along with the current arrival rate of requests, this distinction would be difficult using the current processing rate criteria presented in figure 6.39.

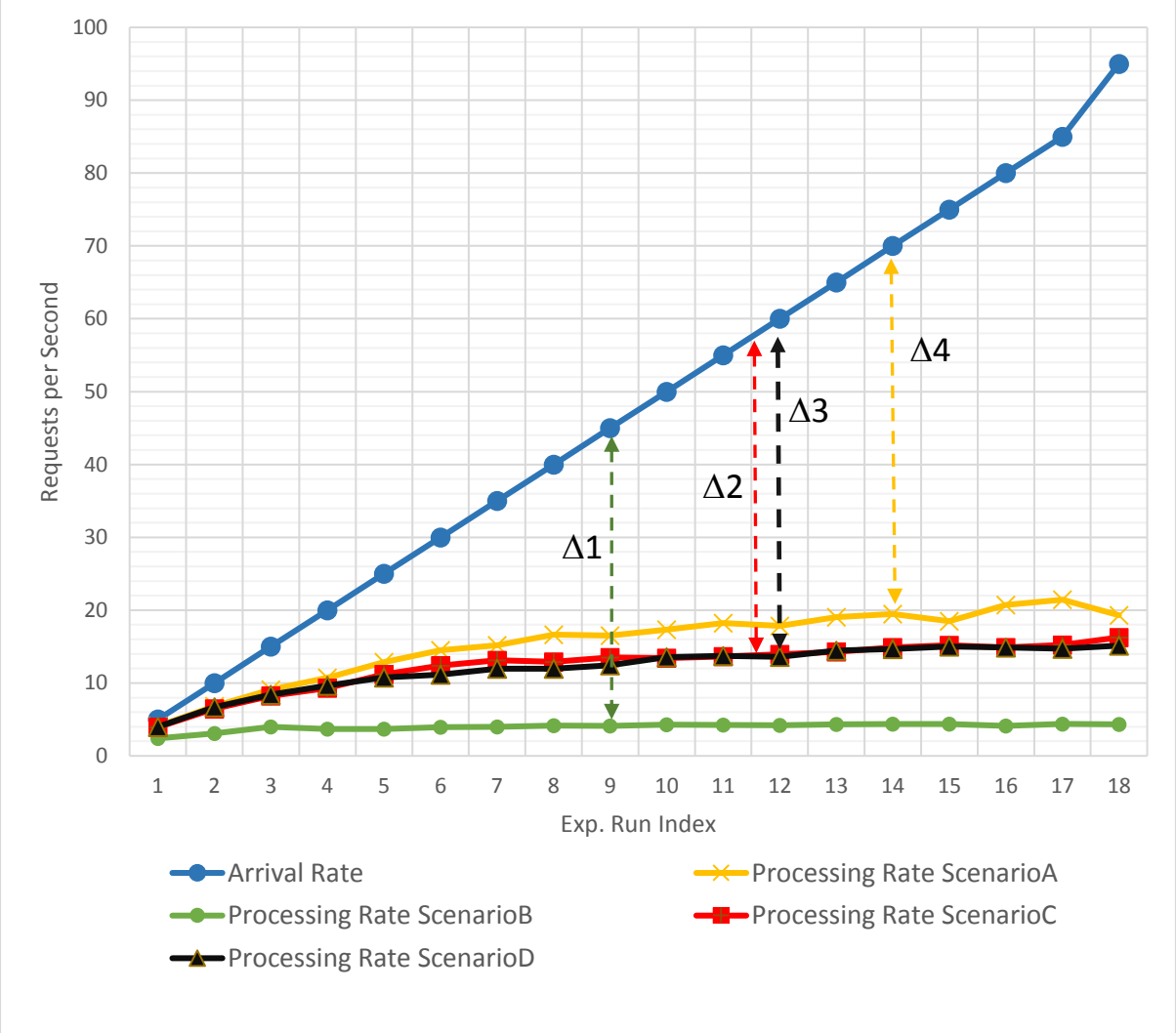


Figure 6.39 Processing Rate Benchmarking for all Scenarios

A new performance measure is proposed that link between the number of served users, average test running time, and processing rate in one performance metric called "System Requests Processing Durability (SRPD)". The proposed performance metric gives weight for both the processing rate (as an indication of system scalability and responsiveness) and the total system running time (as an indication of system availability and reliability),and the number of active core IMS entities (as an indication of overall system utilization and efficiency). The new performance metric is given by equation (6.1).

$$(\boldsymbol{SRPD}) = \frac{\boldsymbol{Total\ Number\ of\ Processed\ Requests}}{\boldsymbol{Total\ Average\ Processing\ Time}} \times (\boldsymbol{Total\ Average\ Processing\ Time})^{\boldsymbol{2}} \times \eta \qquad (6.1)$$

Where $\eta = {}^{\kappa}/_{C}$ represent the system utilization factor; in which κ represent the active cores, and C represent the total number of cores.

Or simply as shown in equation (6.2)

$$(SRPD) = (Total\ Number\ of\ Processed\ Requests) \times (Total\ Average\ Processing\ Time) \times \eta \qquad (6.2)$$

According to this new performance measure, the systems' performance were evaluated again, figure 6.40 shows the performance of the four scenarios. The system utilization for Scenario A was set to 100% as it is not connected to the framework and it has only one IMS subsystem connected directly to the testing system. However, all other scenarios had a system utilization values of 25%, 50%, and 75% for Scenario B, C, and D respectively. According to the new measure, Scenario D has the best performance in terms of processing rate, system availability, scalability and utilization values, followed then in order by Scenario C and B. Scenario A, again, can be used as a baseline performance for the other scenarios as it is not considered part of the proposed framework design. The labels on top of the columns show the enhancement in performance according to the SRPD measure in which up to 333% better performance for scenario D was achieved compared to the baseline performance (Scenario A).

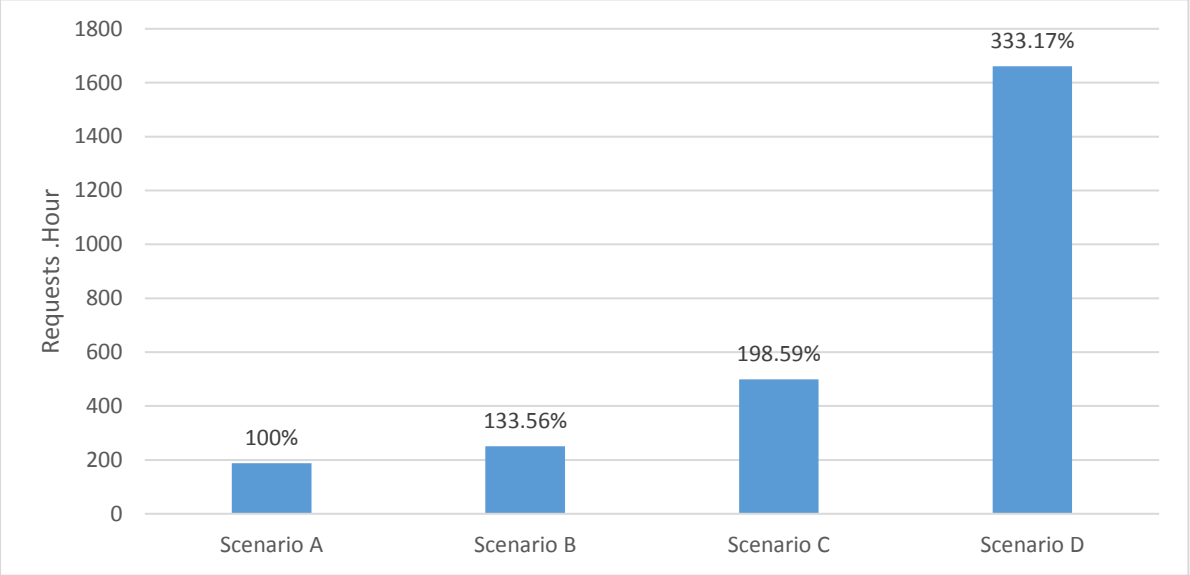


Figure 6.40 System Requests Processing Durability for all scenarios

The comparison between the different scenarios shown in figure 6.40 is useful when a general performance benchmarking is needed. However to let the system run in real time as described in the framework description in the previous chapter, there is a need to overcome the distinction between two scenario's performance estimates. As mentioned previously, scenario C and D had almost the same performance measure. The system need to decide in almost real time the traffic trend and the mapped approximate scenario trend that best describes it. To achieve this goal, a modified form of the SRPD performance measure is used to get the performance estimate in real time within predefined window size that is determined by the resolution unit of the framework.

The new metric is referred as Real Time SRPD (RT- SRPD), in which the performance of the arrived traffic is compared against the same performance measures applied for the SRPD metric described previously (Scalability, Availability, Responsiveness, and Utilization). However, this time the calculation is done within a small window size compared to the entire test running

time for the SRPD. Moreover, the utilization factor as a complement value to get an estimate of system miss-utilization, which is a better representation of the physical meaning of the system performance.

RT- SRPD is shown in formula (6.3). In which the number of requests and the processing time are within a small window size defined by the resolution centre.

$$(RT\_SPRD) = \frac{Number\ of\ Processed\ Requests}{Processing\ Time} \times (Processing\ Time)^2 \times (1 - \eta) \qquad (6.3)$$

This can be rewritten as shown in equation (6.4):

$$(RT\_SPRD) = Number\ of\ Processed\ Requests \times Processing\_Time \times (1 - \eta) \qquad (6.4)$$

The new performance metric was applied over the three scenarios excluding Scenario A and the results are shown in figure 6.41. Scenario A was excluded as it is not using the framework and there is no mechanism to get the requests within predefined window size. As shown in the figure, a clear distinction between the three scenarios is achieved, especially for Scenario C and D. the new thresholds are now simply compared against "zero" as the base line performance reference in this case.

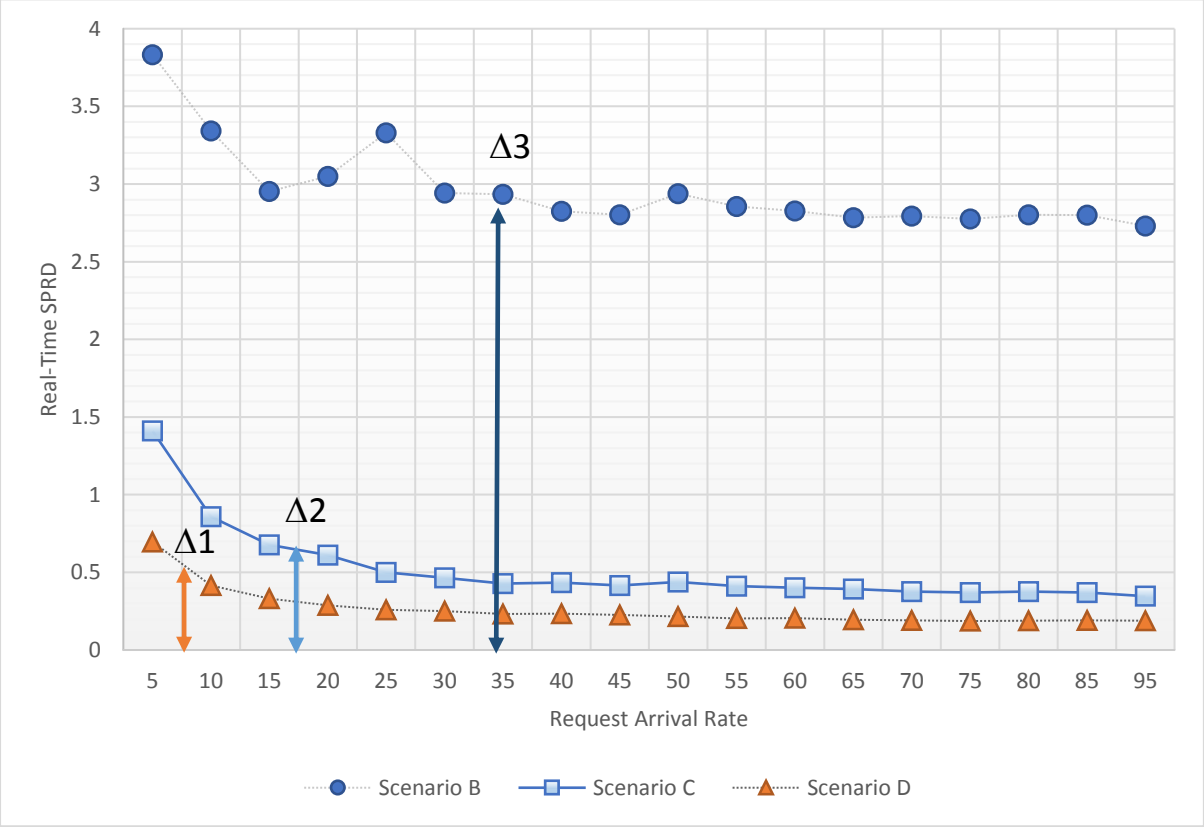


Figure 6.41 Real Time SRPD values vs Arrival Rate

The distinction between different scenarios is one of the core functionalities of the framework, this will help the system to expect the performance based on the current system state and current traffic conditions. To check the validity of the proposed equation in analysing the traffic in real

time, a real time traffic logs was taken at a different random samples and plotted in a scatter plot as shown in figure 6.42.

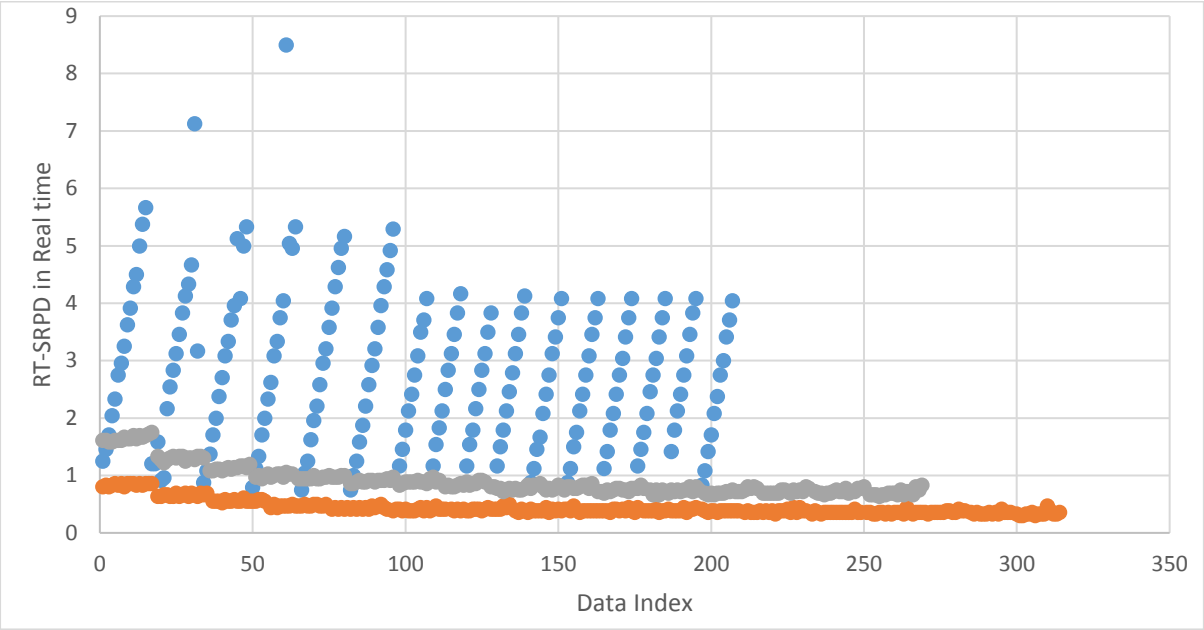


Figure 6.42 RT-SRPD in Real time scatter plot

The scatter shows a distinction and interference intervals. The worst interference of samples was for the first 100 samples of the three scenarios which roughly represent the lower arrival rate of the requests. To get a better view of the interference, figure 6.43 shows the scatter plot again with focus over the first 100 samples of each scenario. This interference in RT- SRPD is due to the large variance value of Scenario A, remember that the distinction between the three scenarios shown in figure 6.42 was based on averaged values for the entire scenario run sample.

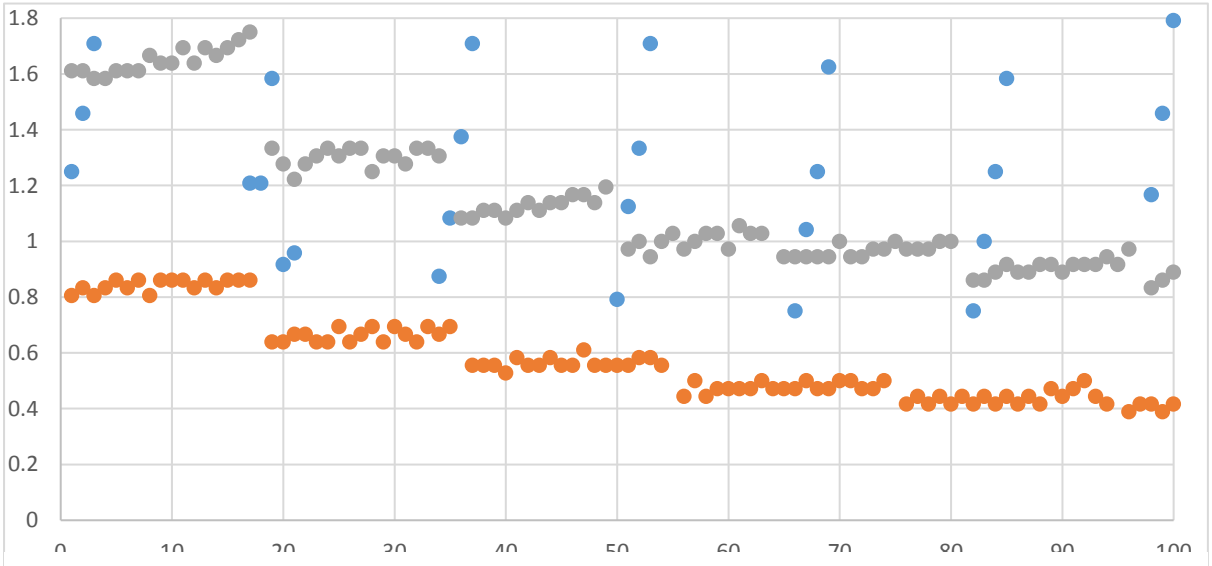


Figure 6.43 Zoomed view of RT-SRPD in Real time scatter plot

This interference can be minimized by keeping track of the SUT state (to distinguish between the three scenarios), and by knowing the inter-arrival rate of the system (λ) to localize the analysis with minimum scattering, and finally by getting the performance trend and the expected future values (by setting a threshold to measure the deviation from the expected average value). The aforementioned three steps will be clarified more next.

The arrival rate is calculated by the intermediate system continuously as described in the framework design chapter. For the test purpose, the average arrival rate of requests is also set by the testing system during the experiment run time.

To get the system state, number of active IMS core entities (κ) need to be known at all times, this can be done by using one of the following two approaches:

1- Get the SUT state by explicit signalling mechanism. The SUT need to notify the intermediate system of its most recent update in real time. The intermediate system will use the feedback information to calculate the system efficiency factor (η) and current RT_ SRPD value. Following this approach, there is no need to calculate κ as it was provided directly by the SUT. This approach is adopted as it can be easily done via the already implemented feedback channel signalling mechanism described in the framework signalling.

2- By reversing the RT- SRPD formula to get the number of active IMS core components κ out of the other variables as shown in formula (6.5). To apply this formula, there is a need to get the number of processed requests and the time taken to process it, this can be easily measured according to the framework functionality described before. However, RT_ SRPD is unknown as it is calculated initially based on the system efficiency vale (η), to get this value the intermediate system (in which the analysis will be performed) need to be aware of the state of the SUT to get the active number of core IMS entities in real time The formula uses average RT_ SRPD values according to previous training set as shown in table 6.5. κ_{ij} then can be calculated as shown in equation 6.5, from which the closest integer value of κ_{ij} is accepted.

$$\kappa_{ij} = C \times \left(1 - \frac{RT\_SPRD_{ij}}{No.Processed\ Requests \times Processing\_Time}\right) \qquad (6.5)$$

$$\forall\, i \in \{1,2,3\}$$

$$\forall\, j \in \{1,2,3,\ldots\ldots,18\}$$

where $RT_{SPRD_{ij}}$ *is the average RTSPRD value of ith Scenario and jth arrival Rate*

Table 6.5 RT-SRPD average values

i / j	Scenario B	Scenario C	Scenario D
	1.00	2.00	3.00
2.00	3.34	0.86	0.42
3.00	2.95	0.68	0.33
4.00	3.05	0.61	0.29
5.00	3.33	0.50	0.26
6.00	2.94	0.46	0.25
7.00	2.93	0.43	0.23
8.00	2.83	0.43	0.23
9.00	2.80	0.42	0.23
10.00	2.94	0.44	0.22
11.00	2.86	0.41	0.20
12.00	2.83	0.40	0.21
13.00	2.78	0.39	0.20
14.00	2.79	0.38	0.19
15.00	2.78	0.37	0.19
16.00	2.80	0.38	0.19
17.00	2.80	0.37	0.19
18.00	2.73	0.35	0.19

Finally the performance trend is determined by comparing actual measured RT- SRPD (RT_SRPD_{new}) against the average RT- SRPD (RT_SRPD_{ij}) and measure the deviation from the average according to equation (6.6)

$$\delta_{ij} = \frac{RT\_SPRD_{new} - RT\_SPRD_{ij}}{RT\_SPRD_{ij}} \qquad (6.6)$$

According to the δ_{ij} value, it will be decided when a jump into next state is needed.

6.7 SCENARIO (F) RESULTS

In this scenario the feedback mechanism is applied and the system operates in a closed loop operation mode. This scenario is fully automated to send the requests to the SUT according to the feedback channel values as described before. The main advantage of this scenario is exploiting the system resources without the need for waiting till the system degrades. This scenario merges between Scenario B, and C by starting with one IMS core entity (similar to the way done in Scenario B), then according to the feedback received send to the other two fresh IMS cores the future requests (similar to the way done in scenario C). This approaches saves the processing power while maximising the system utilisation.

The MATLAB server (Intermediate System) was able to process the traffic batches in real time and analyse the traffic by fitting it as a normal distribution and get the statistics needed to check the proximity to the threshold value. Figures 6.44 shows sample processing of arrived traffic in real time.

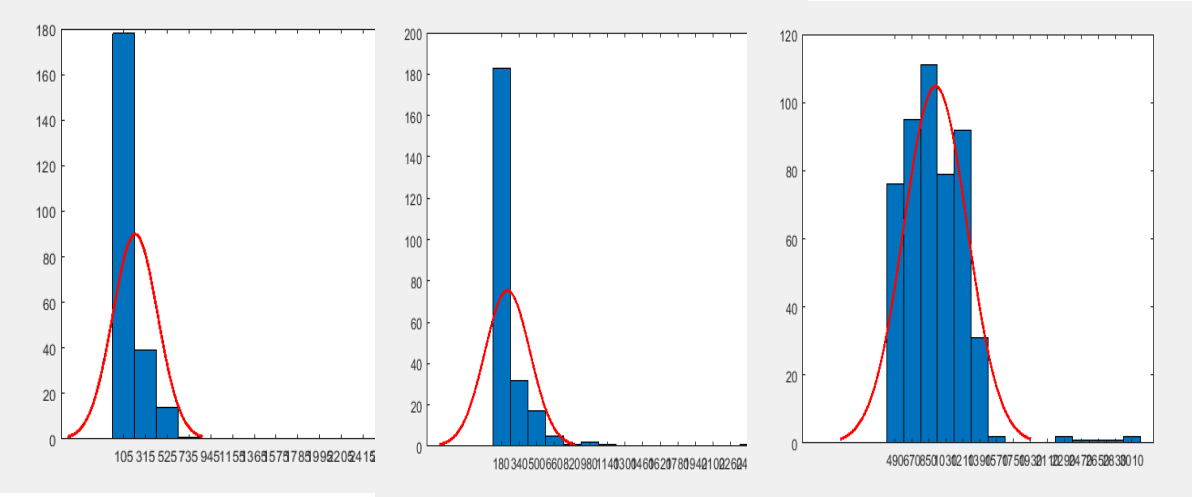


Figure 6.44 Fitted histogram distribution in real time

The total number of served users and the total average requests time is shown in figure 6.45

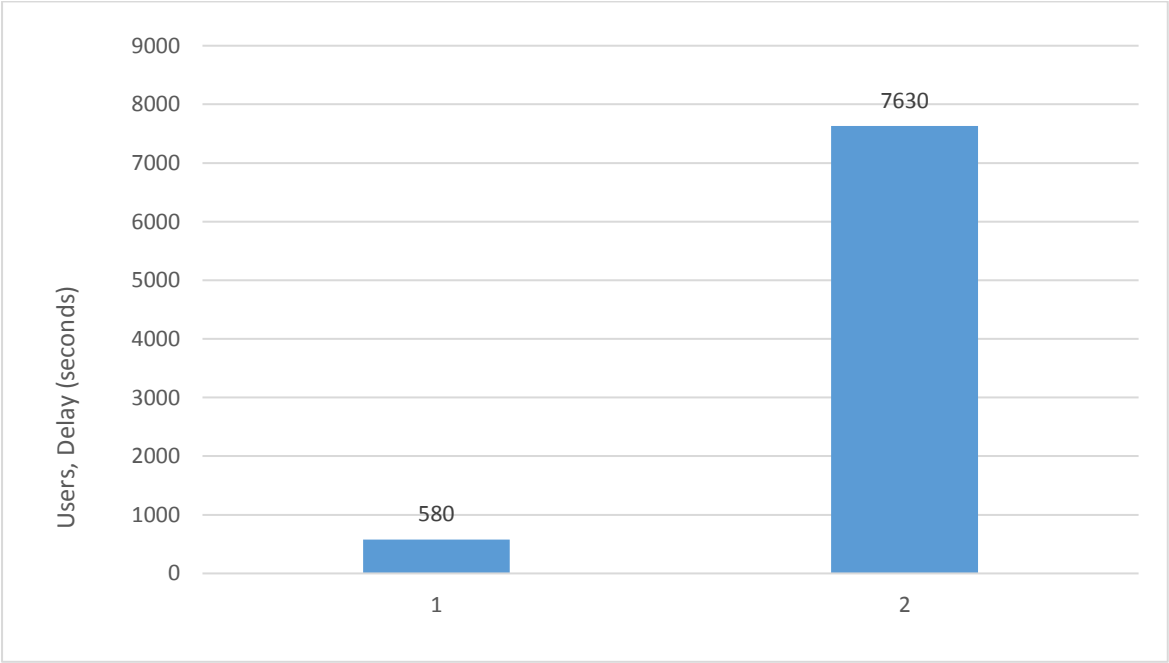


Figure 6.45 Total number of served users and the average requests time

And therefore, the SRPD value = 460 (with an average efficiency of 0.375), which is 255% improved performance compared to scenario A according the SRPD performance measure.

6.8 SUMMERY

Various studies have resulted in many suggestions for enhancing the performance of large-scale systems, such as Mission Critical Communication Systems (MCCSs). However, few have modelled and evaluated the performance of such system in a way that targets overall system performance in real time. Moreover, it is not enough to define the Key Performance Indicators (KPIs) for a system without using them for system performance measurement and performance evaluation. The Session Initiation Protocol (SIP) and IP Multimedia Subsystem (IMS) both have a set of KPIs, such as the registration process delay, that can be used to measure and thus optimize overall system performance. This chapter articulates different options for system simulation and evaluation. The registration process performance affects and reflects the overall system performance. The chapter shows how the registration process delay and the overall system scalability are impacted negatively by system overload.

It was shown from the preliminary simulation results that the number of failed calls *initiation* processes had been increased with the increased number of call pairs, the results focus on the call setup time and the related LTE performance metrics. The optimum number of initiated calls for each pair of calls falls between 150 and 180 for 30 minutes of simulation time with a uniform based distribution system for calls initiation. Moreover, the average number of packets dropped starts between 1 and 3 packets/sec for the single pair scenario and increases to between 6 and 18 packets/sec for the four pairs scenario.

Based on the preliminary experimental results, the RRD with only 100 users each sending one registration request at a time in sequential order was calculated. In which, 90% of Registration requests need less than 40 ms to be completed which meets the requirements of mission critical applications and real-time services. The highest frequency of the registration trials needs on average 20 ms to be completed. This is considered the best-case scenario and was used as a benchmark for the other scenarios in order to enable comparison of both the RRD time and the percentage of trials that finish at certain time threshold. Based on the results, it can be seen that the RRps increases exponentially as the number of users increases up to a limit (of 1,100 users) before beginning to decrease, leading eventually to system degradation and failure.

According to the new defined performance measure, Scenario D (with three active IMS core components) has the best performance in terms of processing rate, system availability, and scalability and utilization values, followed then in order by Scenario C (two core IMS) and B (one core IMS). Scenario A (no intermediate system and one core IMS), again, can be used as a baseline performance for the other scenarios as it is not considered part of the proposed framework design. The labels on top of the columns show the enhancement in performance according to the SRPD measure in which up to 333% better performance for scenario D was achieved compared to the baseline performance (Scenario A).

The deployed solution proved efficiency in increasing the overall system capacity and reliability. The increased number of served user rate indicates that the new framework is able to absorb the sudden increase of user demands due to mission critical situation. System capacity was an indication of the system's ability to be deployed in the core network with thousands of users trying to access the system in a short period of time. The real time sensing mechanism proved that it is able to distinguish between different traffic types and user needs and self-configure the system according to the received load.

Chapter 7: CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

The thesis conclusions can be summarised as follows:

- Mission Critical Communication Systems that are designed to be used with the latest generation of multimedia services are crucial for system users. In order to determine the set of requirements that need to be hardcoded into such systems, a clear distinction between mission-critical and non-mission-critical systems is required. Moreover, the users of and services provided by such systems are very different to those of current mobile commercial communication systems. This implies that there is a set of challenges and requirements that need to be addressed in order to facilitate migration from current systems to those being now being proposed.
- Many challenges need to be addressed and articulated with respect to current commercial mobile communication systems. The main goal is migrating to a generic system that is better in terms of performance than current dedicated mission critical communication systems and at the same time more scalable to serve increasing number of commercial mobile users. The performance of SIP and IMS, as a core part of the generic proposed system, has significant implication on the system's overall performance. The registration process, as an example, is considered a good way to measure the system response time as most of the system entities are involved during the registration process. Similar to the registration, there are a set of KPIs for both the IMS and SIP that can be considered as a reference points to measure the system performance at different layers and abstracts of the end-to-end system. The challenges and requirements that need to be addressed satisfied in any proposed alternative have also been articulated.
- There is a need for exploring the current deployments and research effort in the literature that are related to the SIP and IMS performance to better understand and identify the needed enhancements in SIP and IMS performance, and to get better understanding of the gaps in current research trends. Literature was classified into emerging trends of SIP and IMS performance and the intended research contribution was identified in this thesis. According to the literature, the scalability was one of the major drawbacks of SIP servers and IMS subsystems.

- This thesis implemented a systematic methodology which helped in identifying the bottleneck point for different signalling types. According to the statistics collected for registration and invitation signalling, both the access and core networks experience heavy signalling overhead for serving one user. To have a scalable solution as suggested in the previous chapter, an analysis for possible bottleneck joints in the network. Serving large amount of user may fail part of the network entities. To achieve this goal, a set of tools was used in systematic way to design a framework and set the basis of the evaluation process that will be presented in the next two chapters.

- After defining the framework design methodology, which was described in a systematic way in this thesis, the framework entities was identified and explained, and it was concluded that precise functionality description is crucial to set the evaluation baselines in the following steps. The system performance was identified by a set of flowcharts and algorithms that will be referred to in the later chapters whenever a specific scenario is implemented. The different SUT implementations are described, and it was found that different implementations create differences in performances for different scenarios.

- It was found that there are various studies resulted in many suggestions for enhancing the performance of large-scale systems, such as Mission Critical Communication Systems (MCCSs). However, few have modelled and evaluated the performance of such system in a way that targets overall system performance in real time. Moreover, it is not enough to define the Key Performance Indicators (KPIs) for a system without using them for system performance measurement and performance evaluation. The Session Initiation Protocol (SIP) and IP Multimedia Subsystem (IMS) both have a set of KPIs, such as the registration process delay, that can be used to measure and thus optimize overall system performance. The registration process performance affects and reflects the overall system performance. It was proven that the registration process delay and the overall system scalability are impacted negatively by system overload.

- 5G communications is the new technology that will integrate multiple access technology in to one integrated solution adopted by all vendors and manufacturers. End users and devices will be able to communicate seamlessly with fewer restrictions and more options compared to older technologies Scalability is among the challenges that limit the exploitation of the full capabilities of the current technologies. Many recent studies have tried to overcome the scalability challenge associated with the 5G standards set.

- It was found that the performance of the IP Multimedia Subsystem (IMS) registration process is a key consideration for mission critical systems and is particularly important in large-scale systems where thousands or even millions of users may seek to access the system in disaster scenarios.

- The thesis presents an evaluation of IMS and Session Initiation Protocol (SIP) performance metrics and Key Performance Indicators (KPIs) which are important measure for system benchmarking and performance enhancement. Moreover, it articulates a proposed study that will seek to address some of the challenges identified.

- Based on the simulation results presented in this thesis, it is clear that there is increasing delay in the call setup if LTE communication system was used. This delay is increasing with the number of served clients, which indicates that the Delay requirement or the maximum number of users that can be served at a time may not meet the mission critical service requirements. Hence, the need for decreasing the gap of call setup delay for commercial broadband systems compared with other dedicated mission-critical communications systems is of great importance and considered one of the main challenges for mission-critical communications. This means that there is a need for a new mechanism with minimum possible overhead to minimize access delay by exploring the LTE and IMS domains in addition to the interfaces between LTE and IMS and the interface between LTE and User Element.

- Various studies have resulted in many suggestions for enhancing the performance of large-scale systems, such as Mission Critical Communication Systems (MCCSs). However, few have modelled and evaluated the performance of such system in a way that targets overall system performance in real time. Moreover, it is not enough to define the Key Performance Indicators (KPIs) for a system without using them for system performance measurement and performance evaluation. The Session Initiation Protocol (SIP) and IP Multimedia Subsystem (IMS) both have a set of KPIs, such as the registration process delay, that can be used to measure and thus optimize overall system performance. This chapter articulates different options for system simulation and evaluation. The registration process performance affects and reflects the overall system performance. The chapter shows how the registration process delay and the overall system scalability are impacted negatively by system overload.

- It was shown from the preliminary simulation results that the number of failed calls *initiation* processes had been increased with the increased number of call pairs, the results focus on the call setup time and the related LTE performance metrics. The optimum number of initiated calls for each pair of calls falls between 150 and 180 for 30 minutes of simulation time with a uniform based distribution system for calls initiation. Moreover, the average number of packets dropped starts between 1 and 3 packets/sec for the single pair scenario and increases to between 6 and 18 packets/sec for the four pairs scenario.

- Based on the preliminary experimental results, the RRD with only 100 users each sending one registration request at a time in sequential order was calculated. In which, 90% of Registration requests need less than 40 ms to be completed which meets the requirements of mission critical applications and real-time services. The highest frequency of the registration trials needs on average 20 ms to be completed. This is considered the best-case scenario and was used as a benchmark for the other scenarios in order to enable comparison of both the RRD time and the percentage of trials that

finish at certain time threshold. Based on the results, it can be seen that the RRps increases exponentially as the number of users increases up to a limit (of 1,100 users) before beginning to decrease, leading eventually to system degradation and failure.

- According to the new defined performance measure, Scenario D (with three active IMS core components) has the best performance in terms of processing rate, system availability, and scalability and utilization values, followed then in order by Scenario C (two core IMS) and B (one core IMS). Scenario A (no intermediate system and one core IMS), again, can be used as a baseline performance for the other scenarios as it is not considered part of the proposed framework design. The labels on top of the columns show the enhancement in performance according to the SRPD measure in which up to 333% better performance for scenario D was achieved compared to the baseline performance (Scenario A).

- The deployed solution proved efficiency in increasing the overall system capacity and reliability. The increased number of served user rate indicates that the new framework is able to absorb the sudden increase of user demands due to mission critical situation. System capacity was an indication of the system's ability to be deployed in the core network with thousands of users trying to access the system in a short period of time. The real time sensing mechanism proved that it is able to distinguish between different traffic types and user needs and self-configure the system according to the received load.

7.2 FUTURE WORK

Scientific research in general never stop at certain limits, in the future it is planned to continue and build on the results that was found in this thesis. We can summarise the future work as follows:

1) The simulations in this thesis was implemented with static mobile nodes, the positions of the nodes are fixed. Hence, there is no handoff added complexity for the nodes moving between two cell domains. By mobility, it is not meant only having a moving nodes, but also a dynamic topology that supports handoff mechanisms between the subscriber stations and different base stations. The horizontal handoff is managed via the same access technology using the already standardised implemented handoff mechanisms. However, the vertical handoff mechanisms, in which the user may use different access technologies while moving around, is more challenging as it need to be standardised. It is important for mission critical users, especially in public protection and disaster recovery domain, to be able to use different access domains with seamless handover mechanisms while jumping from one access technology to the other. Though mobility was not targeted in this research, the main techniques implemented for the contribution is mainly deployed in the core network and is considered valid regardless of the implemented access technology. In other words, what applies for the LTE access and EPC signalling analysis, can be similarly applied over other future access technologies using the same applied principles followed in this research,

2) Testing different communication scenarios for an end-to-end connectivity over LTE communication system is needed. For such dynamic topology, the need for measuring the overall performance of the system in terms of SIP signalling and data streaming delay is crucial.

3) Test more scenarios using more powerful computational resources, this will enhance the performance and provide a guideline for the researchers and market operators to understand the optimum operation scenario for certain operational conditions.

4) Increase the number of registered users to exceed the limit of the registered users in the system database which was set to 10000 users. Following the increased computational resources described before.

5) Explore more the potential of using the queuing theory, machine learning, and Artificial Intelligence in controlling the network resources and building a more robust mission critical systems.

REFERENCES

3GPP TS 24.484 version 13.4.0 Release 13, “LTE: Mission Critical Services (MCS) configuration management; Protocol specification“April 2017

3GPP TS 33.180 version 14.1.0 Release 14, “LTE; Security of the mission critical service” October 2017.

3GPP TS 22.281 version 14.4.1 Release 14, “LTE; Mission Critical Video over LTE” April 2018.

3GPP TS 23.280 version 14.1.0 Release 14, “LTE; Common functional architecture to support mission critical services; Stage 2” May 2017.

3GPP TS 23.379 version 14.1.0 Release 14, “LTE; Functional architecture and information flows to support Mission Critical Push To Talk (MCPTT); Stage 2” May 2017.

3GPP TS 23.283 version 15.1.0 Release 15, “LTE; Mission Critical Communication Interworking with Land Mobile Radio Systems” July, 2018.

3GPP TS 23.228, “Service Requirements for the Internet Protocol (IP) Multimedia Core Network Subsystem (IMS), Stage 1.”2005.

3rd Generation Partnership Project; Technical Specification Group Services and System Aspects, Performance Management (PM); Performance measurements; IP Multimedia Subsystems (IMS) (Release 11). 3GPP TS 32.409 (v 11.4.0) September 2012.

3rd Generation Partnership Project; Technical Specification Group Services and System Aspects. Feasibility study on IP Multimedia Subsystems (IMS) evolution (Release 11). 3GPP TR23.812 (v12.1.0) June 2013.

3GPP (2005). Service Requirements for the Internet Protocol (IP) Multimedia Core Network Subsystem (IMS). Technical Specification Group Services and System Aspects, Third Generation Partnership Project.

3GPP (2006). IP Multimedia Subsystem (IMS), Stage 2. TS 23.228 v8.2.0. T. S. G. S. a. S. Aspects.

3GPP (2006). Signalling flows for the IP multimedia call control based Session Initiation Protocol (SIP) and Session Description Protocol. 3GPP TS 24.228 v5.15.0, Third Generation Partnership Project.

3GPP (2012). Performance Management (PM); Performance measurements; IP Multimedia Subsystems (IMS)(Release 11). . Technical Specification Group Services and System Aspects., September 2012. v 11.4.0).

3GPP (2013). Feasibility study on IP Multimedia Subsystems (IMS) evolution TR23.812. Technical Specification Group Services and System Aspects, 3rd Generation Partnership Project.

Arslan Munir and Ann Gordon-Ross, "SIP-Based IMS Signaling Analysis for WiMAX-3G Interworking Architectures," IEEE Transactions on Mobile Computing, vol. 9, no. 5, pp. 733-750, May 2010.

Al-Doski, L.N., R. Ghimire, and S. Mohan, *IP Multimedia Subsystem: Analysis of Scalability and Integration.* Mobile Communications Handbook, 3rd Edition, 2013: p. 695-712.

Ali, A. A., S. Vassilaras and K. Ntagkounakis (2009). A Comparative Study of Bandwidth Requirements of VoIP Codecs over WiMAX Access Networks. 2009 Third International Conference on Next Generation Mobile Applications, Services and Technologies.

Alshamrani, M., H. Cruickshank, Z. Sun, V. Fami, B. Elmasri and E. Danish (2013). Signaling Performance for SIP over IPv6 Mobile Ad-Hoc Network (MANET). 2013 IEEE International Symposium on Multimedia.

Balachandran, K. Budka, K.C Chu, T.P. Doumi, T.L Kang, J.H., R. Whinnery. Converged Wireless Network Architecture for Homeland Security. Military Communications Conference, IEEE MILCOM2005, Atlantic City, NJ, Oct 2005.

Blom, R., de Bruin, P., Eman, J., Folke, M., Hannu, H., Naslund, M., Synnergren, P. (2008). Public Safety Communication using Commercial Cellular Technology. In International Conference and Exhibition on Next Generation Mobile Applications, Services, and Technologies.

Balachandran, K., Budka, K. C., Chu, T. P., Doumi, T. L., & Kang, J. H. (2006). Mobile responder communication networks for public safety. IEEE Communications Magazine, 44(1), 56–64. http://doi.org/10.1109/MCOM.2006.1580933, Jan 2006

Baldini, G., Karanasios, S., Allen, D., & Vergari, F. (2014). Survey of Wireless Communication Technologies for Public Safety. Communications Surveys & Tutorials, IEEE. http://doi.org/10.1109/SURV.2013.082713.00034

Budka, K. C., Chu, T., Doumi, T. L., Brouwer, W., Lamoureux, P., & Palamara, M. E. (2011). Public safety mission critical voice services over LTE. Bell Labs Technical Journal, 16(3), 133–149. http://doi.org/10.1002/bltj.20526

Baldini, G., et al. (2012) "The Evolution of Public Safety Communications in Europe: the results from the FP7 HELP project", ETSI Reconfigurable Radio Systems Workshop.

Baghdadi, M. and S.V. Azhari. *Improving performance of SIP signaling during overload using capabilities of connection-oriented transport protocol*. in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*. 2013.

Baudoin, C., M. Gineste, C. Emmanuel, P. Gelard and J. Bernard (2015). "Dynamic satellite system QoS architecture integrated with IP Multimedia Subsystem core network." International Journal of Satellite Communications and Networking 33(3): 217-239.

Cortes, M., J.R. Ensor, and J.O. Esteban, *On SIP performance.* Bell Labs Technical Journal, 2004. **9**(3): p. 155-172.

Chen, W.E., S.Y. Cheng, and Y.L. Ciou. *A Study on Effects of Different Access Modes on Database Performance for SIP Server*. in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2014.

C. Davids, V.G., S. Poretsky, *Methodology for Benchmarking Session Initiation Protocol (SIP) Devices: Basic session setup and registration*, in *Benchmarking Methodology Working Group*, I.E.T.F. (IETF), Editor. November 12, 2014.

Creswell, J.W., *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 2009: SAGE Publications.

Doumi, T., Dolan, M. F., Tatesh, S., Casati, A., Tsirtsis, G., Anchan, K., & Flore, D. (2013). LTE for public safety networks. IEEE Communications Magazine, 51(2), 106–112. http://doi.org/10.1109/MCOM.2013.6461193

Daniel Diaz-Sanchez, David Proserpio, Andres Marin-Lopez, Florina Mendoza, Peter Weik " A General IMS Registration Protocol for Wireless Network Interworking"2010

D. Malas and A. Morton, "Basic Telephony SIP End-to-End Performance Metrics," Technical Report RFC 6076, Internet Engineering Task Force (IETF), 2011, URL: http://tools.ietf.org/html/rfc6076.

Dacosta, I., V. Balasubramaniyan, M. Ahamad and P. Traynor (2011). "Improving Authentication Performance of Distributed SIP Proxies." IEEE Transactions on Parallel and Distributed Systems 22(11): 1804-1812.

Desai, U., S. Alagesan, A. Goulart and W. Magnussen (2012). Performance of secure SIP and LoST signaling in a Next Generation 9–1–1 testbed. 2012 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR).

Díaz-Sánchez, D., D. Proserpio, A. Marín-López, F. Almenárez-Mendoza and P. Weik (2009). A General IMS Registration Protocol for Wireless Networks Interworking. Wireless and Mobile Networking, Berlin, Heidelberg, Springer Berlin Heidelberg.

ETSI, Europian Telecommunication Standardization Institution http://www.tandcca.com/about/page/12024) ETSI TS 100 392-2 V3.6.1 (2013-05) (last accessed on July 2016)

ETSI, European Telecommunication Standard Institute, IMS Network Testing (INT); IMS/NGN Performance Benchmark. ETSI TS 186 008 November 2012.

ETSI, *IMS Network Testing (INT); IMS/NGN Performance Benchmark;ETSI TS 186 008-2 V2.1.1.* 2013(Part 2: Subsystem Configurations and Benchmarks).

ETSI, *IMS Network Testing (INT); IMS/NGN Performance Benchmark; ETSI TS 186 008-1 V1.2.2.* 2013(Part 1: Core Concepts).

ETSI, *Core Network and Interoperability Testing (INT); IMS/NGN Performance Benchmark;ETSI TS 186 008-4 V2.1.1.* 2015(Part 4: Reference Load network quality parameters).

ETSI, *Core Network and Interoperability Testing (INT); IMS/NGN Performance and Robustness Benchmarking; ETSI TS 186 008-3 V2.1.1.* IMS/NGN Performance Benchmark, Part 3:, 2015(Part 3: Traffic Sets and Traffic Profiles). *Recommendation ITU-T Q.543: "Digital exchange performance design objectives".*

ETSI TS 101 563: "Speech and multimedia Transmission Quality (STQ); IMS/PES/VoLTE exchange performance requirements".

ETSI TS 183 036: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); ISDN/SIP interworking; Protocol specification".

ETSI TS 183 043: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS-based PSTN/ISDN Emulation; Stage 3 specification".

ETSI TS 124 229: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (3GPP TS 24.229)".

ETSI, *Methods for Testing and Specification (MTS); Automated Interoperability Testing; Methodology and Framework*, in *ETSI EG 202 810*. 2010, ETSI.

ETSI, *Methods for Testing and Specification (MTS); Automated Interoperability Testing; Summary of ETSI experiences about using automated interoperability testing tools*, in *ETSI TR 102 789*. 2010.

ETSI, *Methods for Testing and Specification (MTS); Automated Interoperability Testing; Specific Architectures*, in *ETSI TR 102 788*. 2010, ETSI.

ETSI, *IMS Network Testing (INT); IMS & EPC Interoperability test descriptions (3GPP Release 10)*, in *ETSI TS 103 029*. 2013, ETSI.

ETSI TS 101 563: "Speech and multimedia Transmission Quality (STQ); IMS/PES/VoLTE Exchange performance requirements".

ETSI TS 124 229: "Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; IP multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3 (3GPP TS 24.229)".

"ETSI TS 183 036: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); ISDN/SIP interworking; Protocol specification".

"ETSI TS 183 043: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); IMS-based PSTN/ISDN Emulation; Stage 3 specification". Recommendation ITU-T Q.543: "Digital exchange performance design objectives".

ETSI (2010). Methods for Testing and Specification (MTS); Automated Interoperability Testing; Methodology and Framework. ETSI EG 202 810, ETSI. V1.1.1.

ETSI (2010). Methods for Testing and Specification (MTS); Automated Interoperability Testing; Specific Architectures. ETSI TR 102 788, ETSI V1.1.1.

ETSI (2010). Methods for Testing and Specification (MTS); Automated Interoperability Testing; Summary of ETSI experiences about using automated interoperability testing tools. ETSI TR 102 789. V1.1.1.

ETSI (2012). IMS Network Testing (INT); IMS/NGN Performance Benchmark. ETSI TS 186 008, European Telecommunication Standard Institute.

ETSI (2013). IMS Network Testing (INT); IMS & EPC Interoperability test descriptions (3GPP Release 10). ETSI TS 103 029, ETSI. V5.1.1.

ETSI (2013). "IMS Network Testing (INT); IMS/NGN Performance Benchmark; ETSI TS 186 008-1 V1.2.2." (Part 1: Core Concepts).

ETSI (2013). "IMS Network Testing (INT); IMS/NGN Performance Benchmark;ETSI TS 186 008-2 V2.1.1." (Part 2: Subsystem Configurations and Benchmarks).

ETSI (2015). "Core Network and Interoperability Testing (INT); IMS/NGN Performance and Robustness Benchmarking; ETSI TS 186 008-3 V2.1.1." IMS/NGN Performance Benchmark, Part 3:(Part 3: Traffic Sets and Traffic Profiles).

ETSI (2015). "Core Network and Interoperability Testing (INT); IMS/NGN Performance Benchmark;ETSI TS 186 008-4 V2.1.1." (Part 4: Reference Load network quality parameters).

Ferrús, R., Sallent, O., Baldini, G., & Goratti, L. (2013). LTE: The technology driver for future public safety communications. IEEE Communications Magazine, 51(10), 154–161. http://doi.org/10.1109/MCOM.2013.6619579

Farahbaksh R., Varposhti M., Movahhedinia N." Transmission Delay reduction in IMS by Re-registration Procedure modification" the second International conference on Next Generation Mobile Applications, Services and Technologies. 2007.

Femminella, M., E. Maccherani, and G. Reali. *Performance management of Java-based SIP application servers*. in *12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*. 2011.

Farahbaksh R., V. M., Movahhedinia N. (2007). Transmission Delay reduction in IMS by Re-registration Procedure modification. the second International conference on Next Generation Mobile Applications, Services and Technologies.

Fokus. (2004). "Open IMS core." from http://www.openimscore.org. (last accessed on 14th May 2018)

Gurbani, V.K., L.J. Jagadeesan, and V.B. Mendiratta, *Characterizing session initiation protocol (SIP) network performance and reliability.* Service Availability, 2005. **3694**: p. 196-211.

Happenhofer, M., C. Egger, and P. Reichl. *Quality of signalling: A new concept for evaluating the performance of non-INVITE SIP transactions*. in *2010 22nd International Teletraffic Congress (ITC 22)*. 2010.

Happenhofer, M. and P. Reichl. *Quality of signaling (QoSg) metrics for evaluating SIP transaction performance*. in *SoftCOM 2010, 18th International Conference on Software, Telecommunications and Computer Networks*. 2010.

Hossain, I. S., S. H. S. Ariffin, N. Fisal, N. S. A. Hassan, L. A. Latiff and C. K. Neng (2011). Performance analysis indoor location tracking framework with SIP on IPv6. 2011 Fourth International Conference on Modeling, Simulation and Applied Optimization.

Husić, J. B., H. Bajrić, E. Neković and S. Baraković (2012). Basic telephony SIP end - to - end performance metrics. 2012 IX International Symposium on Telecommunications (BIHTEL).

IPWireless. LTE addressing the needs of the Public Safety Community. 3GPP RAN Workshop on Rel-12 and Onward RWS-120030. June 2012.

ITU-T TR Q-series supplements 51 signaling requirements for IP-QoS (December 2004).

ITU-T (2004). Signalling requirements for IP-QoS TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU. SERIES Q: SWITCHING AND SIGNALLING

James Rankin, Alexandru Costaiche, and Joseph Zeto "Validating VoLTE – A Definitive Guide to Successful Deployments" 1st edition, 2013, IXIA, 915-4020-01.

Jia, Z., Z. Liang, and Y. Dai, *Scalability Evaluation and Optimization of Multi-Core SIP Proxy Server*, in *2008 37th International Conference on Parallel Processing*. 2008. p. 43-50.

James Rankin, A. C., Joseph Zeto, Kathy O'Neil (2014). "Validating VoLTE: A Definitive Guide to Successful Deployments."

Jiang, H., A. Iyengar, E. Nahum, W. Segmuller, A. N. Tantawi and C. P. Wright (2012). "Design, Implementation, and Performance of a Load Balancer for SIP Server Clusters." IEEE/ACM Transactions on Networking 20(4): 1190-1202.

Jiri Hosek, L. N., Vit Novotny, Pavel Masek, Dominik Kovac (2004). Performance Analysis: Impact of Signalling Load over IMS Core on KPIs. Recent Advances in Circuits, systems, Telecommunications and Control. I.-T. TR.

Kulin, M., T. Kazaz, and S. Mrdovic. *SIP server security with TLS: Relative performance evaluation*. in *2012 IX International Symposium on Telecommunications (BIHTEL)*. 2012.

Krishnamurthy, R. and G.N. Rouskas. *On the impact of scheduler settings on the performance of multi-threaded SIP servers*. in *2015 IEEE International Conference on Communications (ICC)*. 2015.

Krishnamurthy, R. and G.N. Rouskas. *Evaluation of SIP proxy server performance: Packet-level measurements and queuing model*. in *2013 IEEE International Conference on Communications (ICC)*. 2013.

Krishnamurthy, R. and G.N. Rouskas. *Performance evaluation of multi-core, multi-threaded SIP proxy servers (SPS)*. in *2016 IEEE International Conference on Communications (ICC)*. 2016.

Kellokoski, J., et al. *Call and messaging performance comparison between IMS and SIP networks*. in *2010 IEEE 4th International Conference on Internet Multimedia Services Architecture and Application*. 2010.

Kleinrock, L., *Queueing Systems, : Theory*. Vol. Vol. 1. 1975, New York: Wiley.

Kellovsky, M. and I. Baronak, *Ip Multimedia Subsystem - Dimensioning of the Home Subscriber Server Database.* Journal of Electrical Engineering-Elektrotechnicky Casopis, 2014. **65**(6): p. 376-380.

Kist, A.A. and R.J. Harris, *Sip signalling delay in 3GPP.* Converged Networking: Data and Real-Time Communications over Ip, 2003. **119**: p. 211-222.

Kueh, V.Y.H., R. Tafazolli, and B.G. Evans, *Performance analysis of session initiation protocol based call set-up over satellite-UMTS network.* Computer Communications, 2005. **28**(12): p. 1416-1427.

Lu, C. (2012). Delay Analysis of Push to Talk over Cellular (PoC) Service Solutions for Public Safety Communications Over LTE Networks Master Thesis, University Politicnica De Barcelona.

Mishra, G., S. Dharmaraja, and S. Kar. *Performance analysis of SIP signaling network using hierarchical modeling*. in *2014 Twentieth National Conference on Communications (NCC)*. 2014.

McGee, A. R., Coutière, M., & Palamara, M. E. (2012). Public safety network security considerations. Bell Labs Technical Journal, 17(3), 79–86. http://doi.org/10.1002/bltj.21559

Mukherjee, S. and C. Beard. *A framework for ultra-reliable low latency mission-critical communication*. in *2017 Wireless Telecommunications Symposium (WTS)*. 2017.

Montagna, S. and M. Pignolo, *Performance Evaluation of Load Control Techniques in SIP Signaling Servers*, in *Third International Conference on Systems (icons 2008)*. 2008. p. 51-56.

Munir, A., & Gordon-Ross, A. (2010). SIP-Based IMS Signaling Analysis for WiMax-3G Interworking Architectures. *Ieee Transactions on Mobile Computing, 9*(5), 733-750. doi:10.1109/TMC.2010.16.

Nader Mir, Sarhan Musa, Heng Gao, Chaitra shivakumar “Performance Analysis of IMS Signaling Multimedia Networks” Information Engineering (IE), Volume 1 Issue 1 2012.

“On SIP performance” Bell Labs Technical Journal, Wiley Subscription Services, Inc., A Wiley Company, 2004.

Over, C. O., & Etworks, L. T. E. N. (2012). D ELAY A NALYSIS OF P USH - TO -T ALK OVER C ELLULAR (P O C) S ERVICE S OLUTIONS FOR P UBLIC S AFETY, 1–48.

“Public Safety mobile broadband and spectrum needs”, Report for the TETRA Association, 8 March 2010, 16395-94, Analysis Mason.

Project 25 System and Standard Definition, TIA, TSB102-A, 1995 (http://www.project25.org) (last accessed on 20th March 2016)

Paper, N.W., *5G for Mission Critical Communication: Achieve ultra-reliability and virtual zero latency*. 2016.

Poretsky, C.D.S., *Terminology for Benchmarking Session Initiation Protocol (SIP) Networking Devices*, in *Request for Comments: 7501*, I.E.T.F. (IETF), Editor. April 2015.

Rolf Blom, Peter de Bruin, Jesper Eman, Mats Folke, et al. Public Safety Communication Using Commercial Cellular Technology. The second International Conference on Next Generation Mobile Applications, Services and Technologies, 2008.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

RAJAGOPAL, N., *Modeling and Performance Prediction of IP Multimedia Subsystem Networks.* Thesis, 2006.

Sanjay Kanti Das. Feasibility study of IP Multimedia Subsystems (IMS) based Push To Talk over Cellular for Public Safety and Security Communications. Master's Thesis Department of Electrical and Communication Engineering. Helsinki University of Technology (HUT), 2006.

Simon Forge, Robert Horvitz and Colin Blackman "Is Commercial Cellular Suitable for Mission Critical Broadband?" ISBN: 978-92-79-38679-4, 2014.

Simic, M. B. (2012). Feasibility of long term evolution (LTE) as technology for public safety. In 2012 20th Telecommunications Forum (TELFOR) (pp. 158–161). http://doi.org/10.1109/TELFOR.2012.6419172

Signalling flows for the IP multimedia call control based Session Initiation Protocol (SIP) and Session Description Protocol; Stage3 (Release 5), 3GPP TS 24.228 v5.15.0, 2006.

Subramanian, S.V. and R. Dutta. *Comparative Study of M/M/1 and M/D/1 Models of a SIP Proxy Server*. in *2008 Australasian Telecommunication Networks and Applications Conference*. 2008.

Subramanian, S.V. and R. Dutta. *Measurements and Analysis of M/M/1 and M/M/c Queuing Models of the SIP Proxy Server*. in *2009 Proceedings of 18th International Conference on Computer Communications and Networks*. 2009.

Subramanian, S.V. and R. Dutta. *Performance and scalability of M/M/c based queuing model of the SIP Proxy Server - a practical approach*. in *2009 Australasian Telecommunication Networks and Applications Conference (ATNAC)*. 2009.

Shen, C., et al., *The Impact of TLS on SIP Server Performance: Measurement and Modeling.* IEEE/ACM Transactions on Networking, 2012. **20**(4): p. 1217-1230.

Spirent, *IMS Procedures and Protocols: The LTE User Equipment Perspective*. 2013. p. 31.

Shen, C., E. Nahum, H. Schulzrinne and C. P. Wright (2012). "The Impact of TLS on SIP Server Performance: Measurement and Modeling." IEEE/ACM Transactions on Networking 20(4): 1217-1230.

Subramanian, S. V. and R. Dutta (2008). Comparative Study of M/M/1 and M/D/1 Models of a SIP Proxy Server. 2008 Australasian Telecommunication Networks and Applications Conference.

Subramanian, S. V. and R. Dutta (2009). Measurements and Analysis of M/M/1 and M/M/c Queuing Models of the SIP Proxy Server. 2009 Proceedings of 18th International Conference on Computer Communications and Networks.

Subramanian, S. V. and R. Dutta (2010). Performance Measurements and Analysis of M/M/c Queuing Model Based SIP Proxy Servers in Local and Wide Area Networks. 2010 International Conference on Advances in Recent Technologies in Communication and Computing.

Terrestrial Trunked Radio (TETRA); Voice plus Data (V+D); Part 2: Air Interface (AI), ETSI, EN 300 392-2 v2.3.10, 2003.

Third Generation Partnership Project, Organization, 3GPP. (2008). 3rd Generation Partnership Project ; Technical Specification Group Radio Access Network. Evolved Universal Terrestrial Radio Access (EUTRA).

TETRA and Critical Communication Association (TCCA) (http://www.tandcca.com/), 2010. (last accessed on 11 July 2017)

TETRA MoU Association. Push To Talk over Cellular (PoC) and Professional Mobile Radio (PMR), TETRA 2004.

Technical Specification Group Services and System Aspects (2006), IP Multimedia Subsystem (IMS), Stage 2, TS 23.228, 3rd Generation Partnership Project TS 23.228 v8.2.0, 2007

Voznak, M. and J. Rozhon (2010). SIP Back to Back User Benchmarking. 2010 6th International Conference on Wireless and Mobile Communications.

Wright, C. P., E. M. Nahum, D. Wood, J. M. Tracey and E. C. Hu (2010). "SIP server performance on multicore systems." IBM Journal of Research and Development 54(1): 7:1-7:12.

ZHU, B., *Analysis of SIP in UMTS IP Multimedia Subsystem*. 2003, University North Carolina.

INDEX

K

L

M

N

O

P

Q

R

S

T

U

V

W

APPENDICES

APPENDIX A: EXPERIMENT SETUP (MATLAB AND VM SCREENSHOTS)

In this appendix a brief description of the test bed experiment will be described. As explained in the thesis, the test be is built using different set of servers that are interconnected to each other. For demonstration purposes only, each server was imaged over a virtual machine running over a powerful workstation. The servers and virtual machines order are described as follows:

VM1: loaded with the traffic generator server along with all needed dependencies in addition to the PhP/SQL server instances. The server has it is own web console accessible by the admin to monitor the HSS databases remotely.

VM2: loaded with the first IMS core component running FOKUS Open IMS core subsystem and equipped with traffic monitoring tools for monitoring and analysis purposes.

VM3: loaded with the second IMS core component running FOKUS Open IMS core subsystem and equipped with traffic monitoring tools for monitoring and analysis purposes.

VM4: loaded with the third IMS core component running FOKUS Open IMS core subsystem and equipped with traffic monitoring tools for monitoring and analysis purposes.

VM5: equipped with MATLAB server that is able to sniff the network traffic via added Simulink model that has all the dependencies needed for this task.

It is important to notice that all servers are running over the virtual machine equivalently to the intended way it was set over the real physical machines in the lab. The switching of the traffic was implemented using more than one virtual switch interconnecting the VMs (thanks to the powerful add-ons of VMWARE workstation version).

It is important to mention as well that both the PhP server and the traffic generator are merged in one entity (one single virtual machine) without loss of intended functionality purpose. Both servers are running separately without any intersections as both are doing two different tasks, this separation was made to save the processing resources.

It worth mentioning as well that the traffic analysis using MATLAB was the core functionality of the intermediate system described in the thesis and both are implemented over MATLAB server. Other secondary functions, such as estimating the load stress, was assigned to the traffic generator server and fed to Matlab server directly. The logical separation of the intermediate system functionality did not change the intended functionality described in the thesis.

Figure A.1 shows a screenshot of the traffic generator (VM1) as described in the thesis trying to send 7000 registered users to the core system.

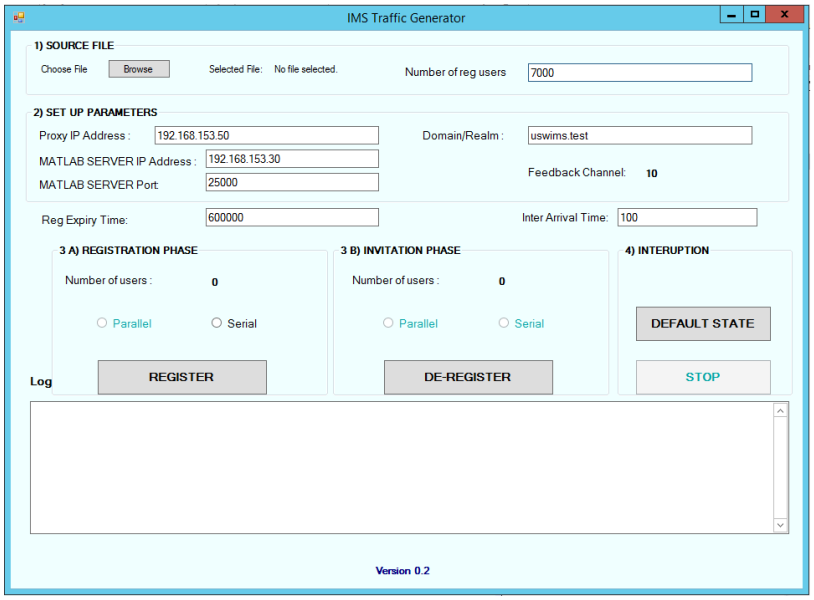


Figure A.5 Packet Generator VM1

Figure A.2 shows the successful transmission of SIP messages as intended along with the header values and IP addresses. 200OK message indicate successful transmission, and 500 error message shows the failure of one of the core IMS entities due to the overload added. The failure will be discussed later in other figures.

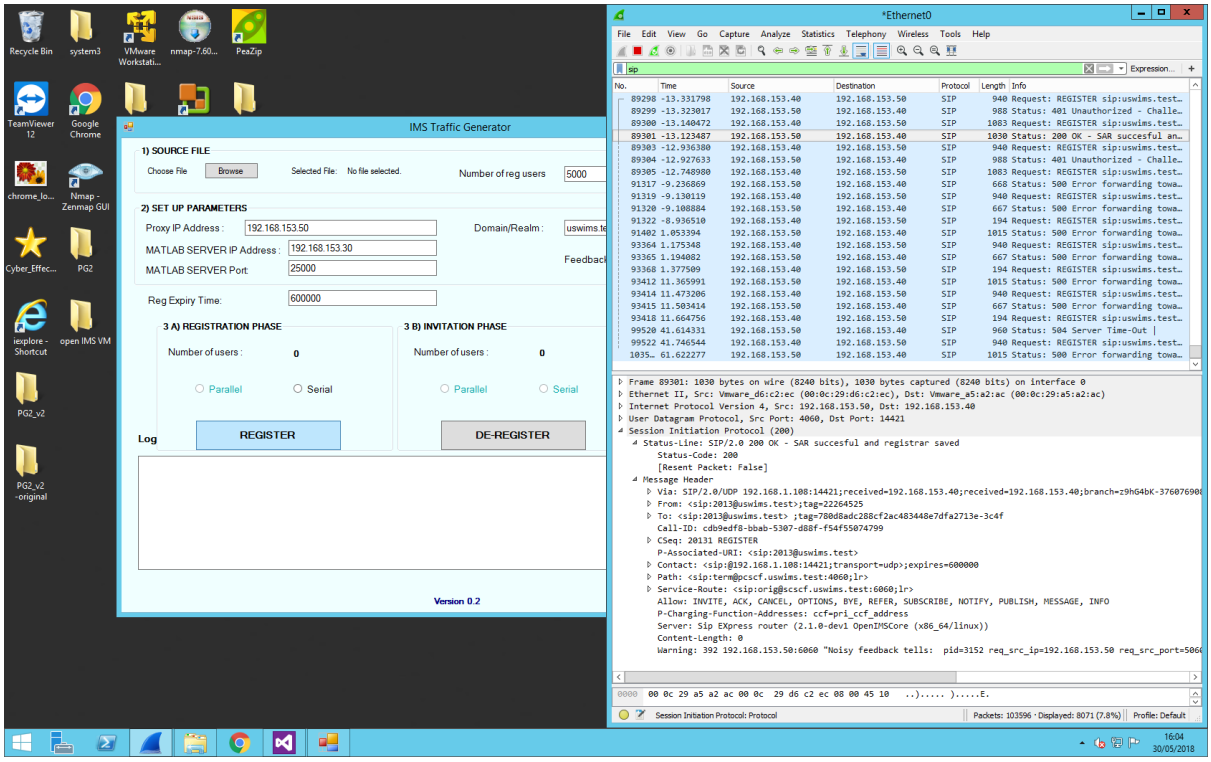


Figure A.6 Traffic received at IMS1 (VM2)

Figure A.3 shows Matlab server (VM5). It shows the listening state (the waiting state until a traffic is received) then it shows the running state (by showing "Listening …….. 3" message) to indicate the there is a traffic received and being analysed in the background.

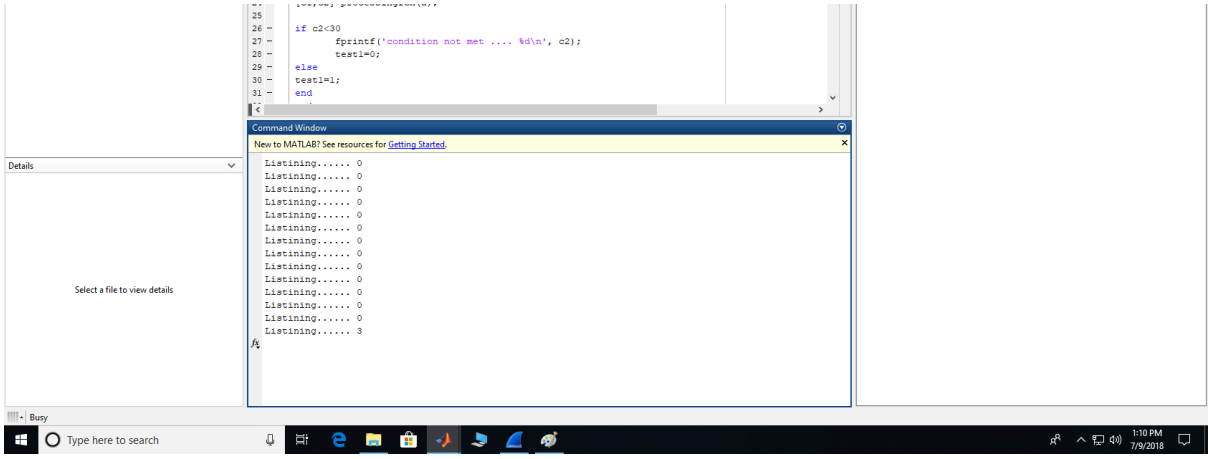


Figure A.7 Matlab Server Console Panel (VM5)

Figure A.4 shows the Simulink model running. The system was able to get the network traffic in real time and pass the loaded delay data into MATLAB workspace for further processing.

Each running cycle (Epoch time) takes 50 seconds which is the adopted window size that is set based on the resolution centre described in the thesis.

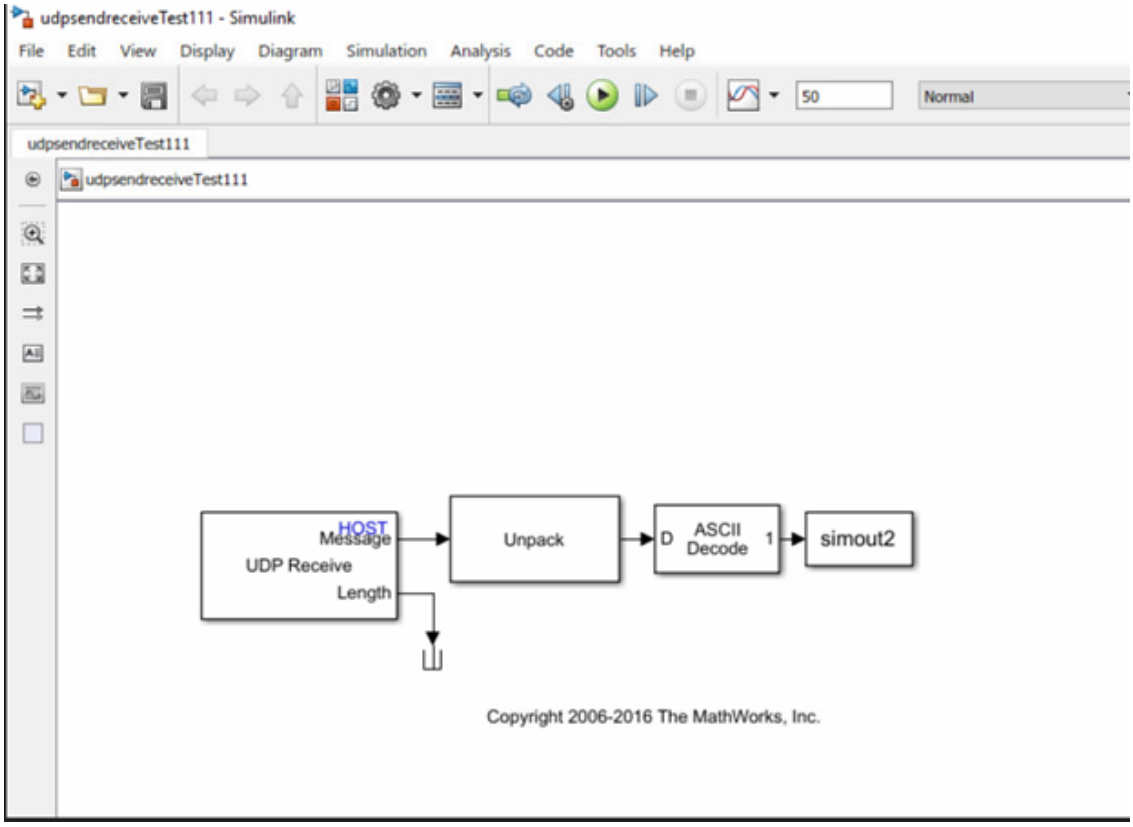


Figure A.8 Simulink Model interface (VM5)

Following the first cycle of analysis described before. Matlab applies the traffic analysis logic as described in the thesis and decide in real time (within the allocated window size) if the system is overloaded or need being overloaded. Figure A.5 shows three running cycles in which the condition of overload testing criteria was not met. The outcome of the function value is returned as well.

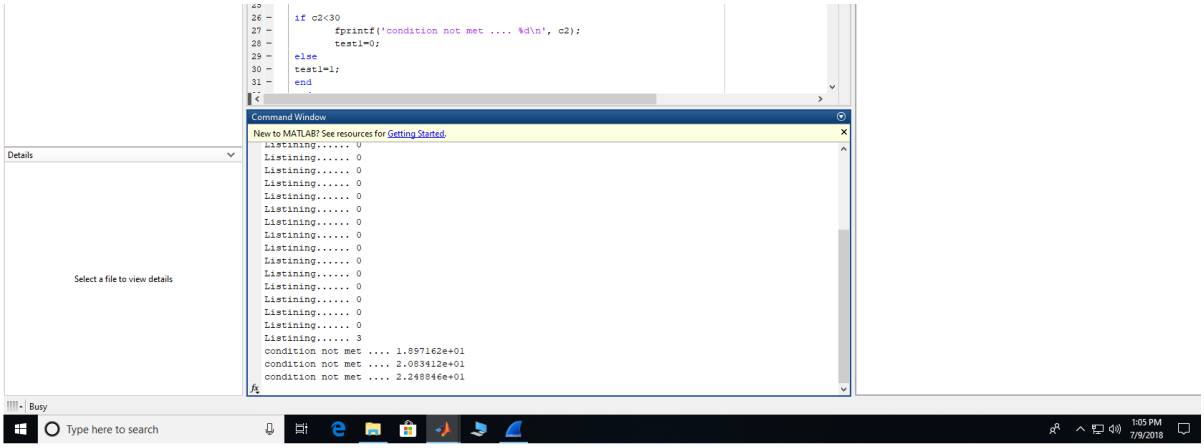


Figure A.9 Listening and Processing cycles

On another instant, figure A.6 shows that the condition was met successfully and a control message over UDP (origination from port number 25000) was sent successfully to the traffic generation server. The message is loaded with the Feedback channel value.

z

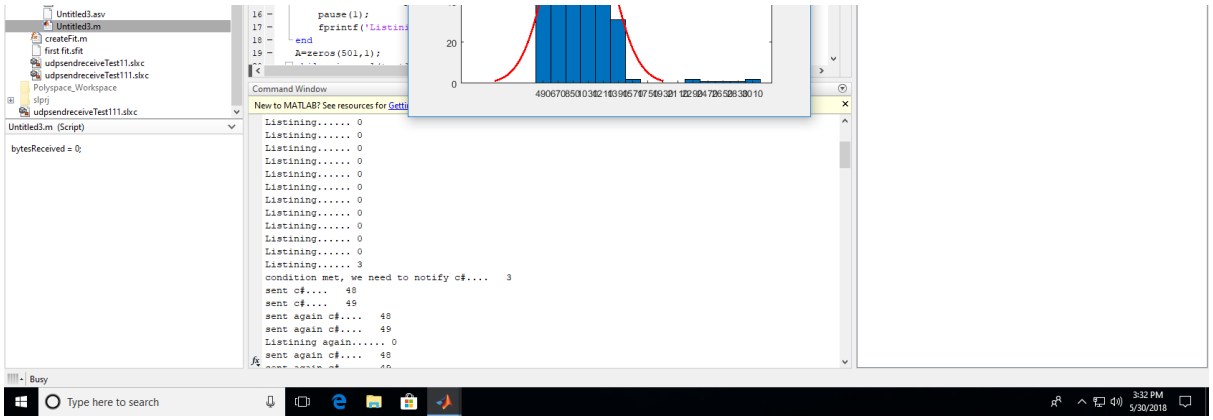


Figure A.10 Feedback Channel control signaling

Following the sent feedback channel, the system continues monitoring and analysis the traffic to ge the new set of data. Figure 7 shows another parsing cycle while in which it shows that the generator responded successfully to the overload state and distribuated the load. The histogram shows that the load per server, and the over all delay was reduced.

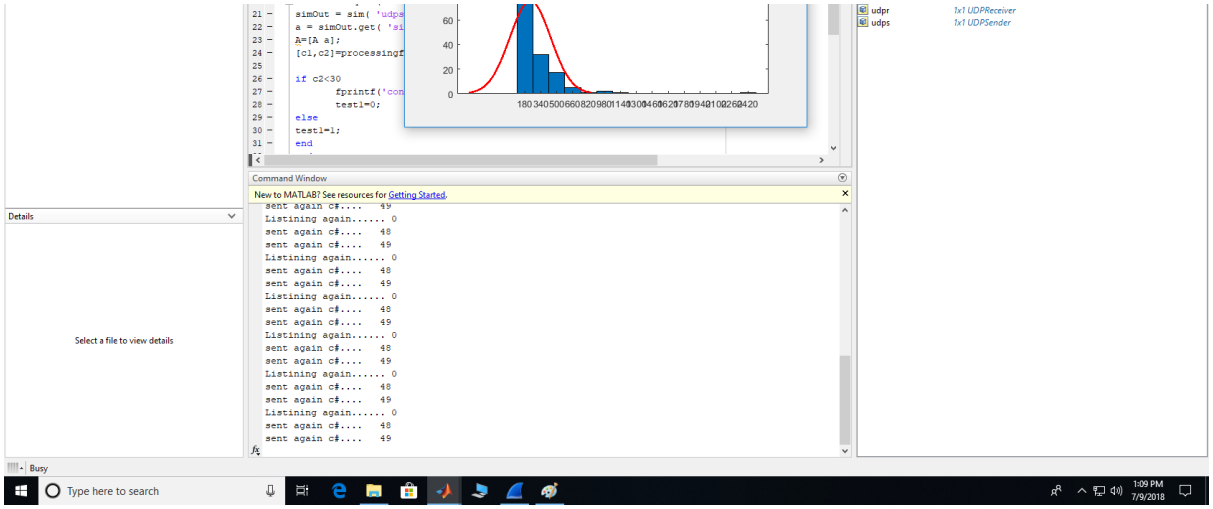


Figure A.7 Reduced load following FBC send request

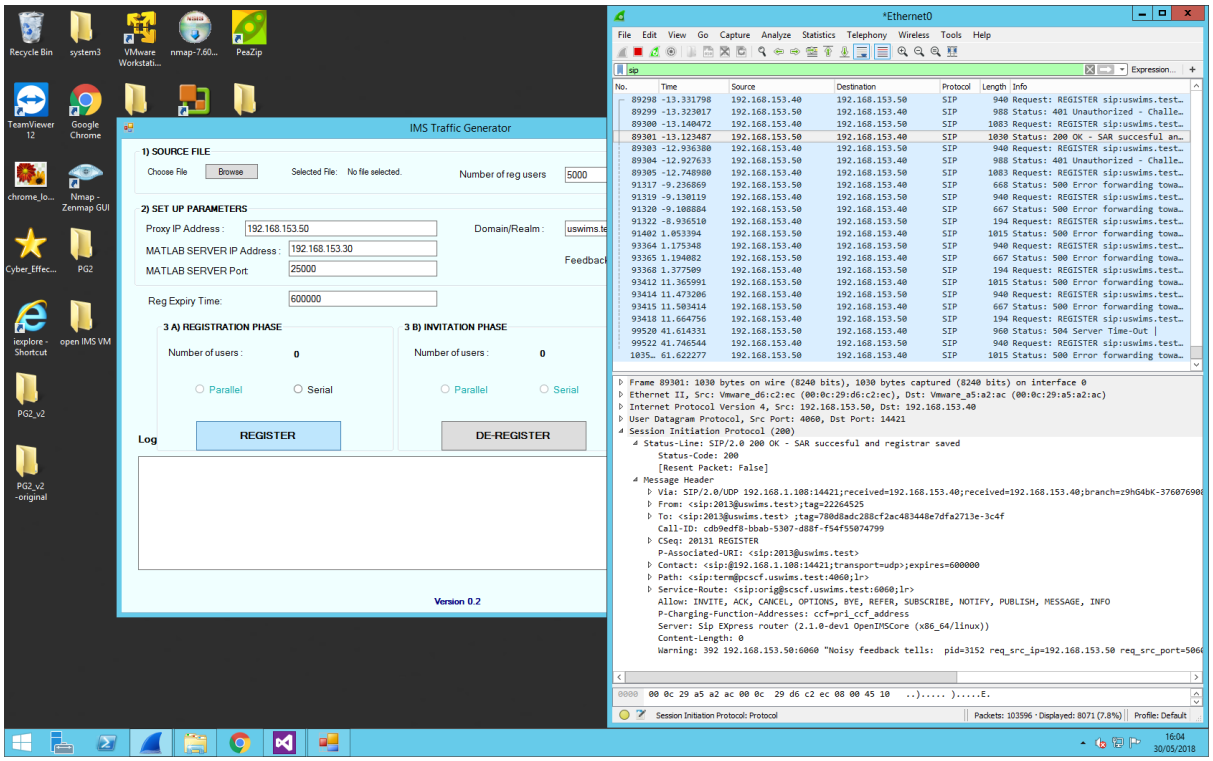


Figure A.8 Reduced load following FBC send request

APPENDIX B: OPNET (RIVERBED) SIMULATION

Riverbed Modeler or what used to be (OPNET) is a Suite of Protocols and Technologies with a Sophisticated Development Environment. By modeling all network types and technologies (including LTE,VoIP, TCP, OSPFv3, MPLS, IPv6, and more), Riverbed Modeler analyzes networks to compare the impact of different technology designs on end-to-end behavior. Modeler lets you test and demonstrate technology designs before production; increase network R&D productivity; develop proprietary wireless protocols and technologies; and evaluate enhancements to standards-based protocols (www.riverbed.com, 2018).

Network designers constantly challenged by the growing complexity of communications protocols and the increasing scale of network deployments and Network R&D is no longer a process that can be conceded to spreadsheets or traditional software. Therefore, In order for Network R&D organizations to innovate, they need robust network simulation software to easily and intuitively model the intricate end-to-end behavior of protocols. The solution must also be able to efficiently analyze the performance of these protocols and technologies in network infrastructure models of realistic scale.

This thesis has presented results based on the OPNET simulations in which the LTE modile was used in top of the basic OPNET simualtions. The main advantage of OPNET is its discrete event simulation cababilities, in which the events are driven in sequential order similar to the whay physical networking systems are operating. The network level, nodes level, and process levels are shoun in figure B.1 below

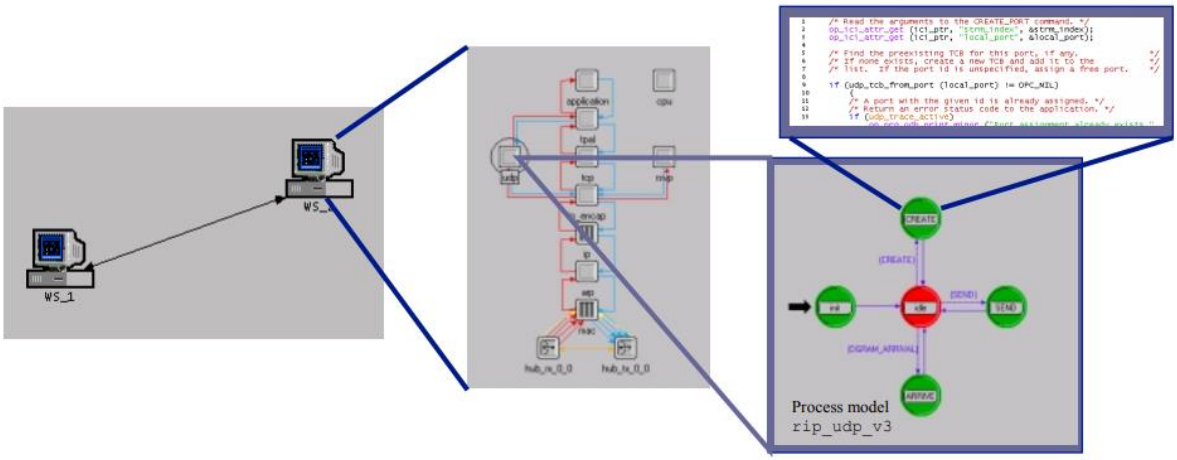


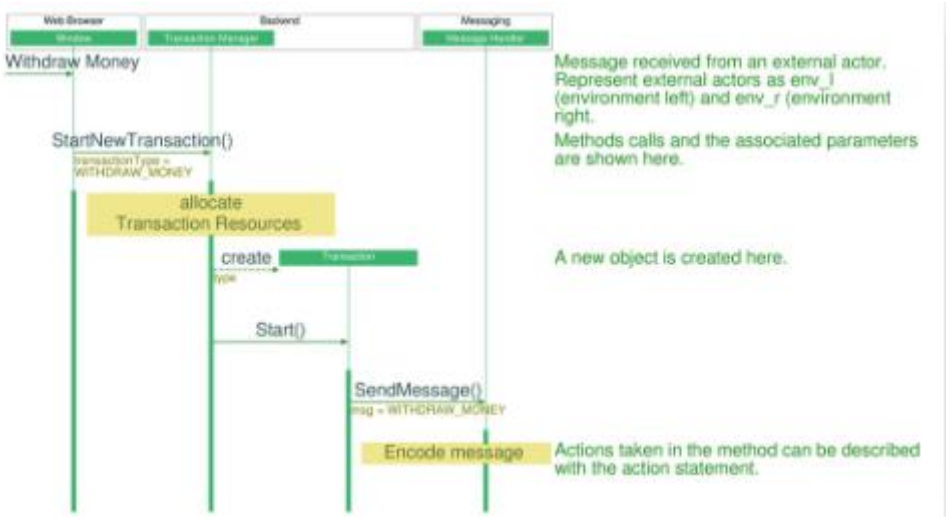
Figure B.1 Reduced load following FBC send request

As described in the thesis, the LTE and SIP modules where used in the OPNET simulation. Although there are some IMS implementations over OPNET in the literature, none of them is mature and a lot are under development. The simulations carried out in the thesis are designed mainly to test the access technology delays and the single SIP server added overhead.

APPENDIX C: EVENT HELIX

In this thesis, Eventhelix, which has a script language, was used to get the flowcharts and sequence diagrams in addition to state transition diagrams. Telecom signaling interactions in general are difficult to document and understand due to their incredible complexity. EventStudio helps by representing the system as a multi-level hierarchy and then letting you generate diagrams at different levels of abstraction. (www.eventlelix.com, 2018)

Messages, timers, resources modeling is built into the FDL modeling language. Multiple scenario modeling support helps with modeling of myriads of success and failure signaling scenarios. EventStudio has built in support for object interaction modeling. It starts with defining the architecture as a multi-level hierarchy. The different levels could be used to represent namespaces and modules. Object interactions are represented as a cascade of method invokes and returns as shown in figure C.1. New and delete of objects modeling is directly supported. EventStudio even analyzes your design to catch memory leaks in scenarios.



This tool was used as well to count the number of hits during the message interaction sequence while travelling between the nodes, the statistics collected were summarised in the tables presented in the bottleneck analysis section in this thesis.